A PHYLOGENOMIC INVESTIGATION INTO THE EVOLUTION AND BIOLOGICAL CHARACTERISTICS OF THE BEIJING LINEAGE FAMILY OF PRINCIPLE GENETIC GROUP 1 MEMBERS OF *MYCOBACTERIUM TUBERCULOSIS*

By

Kabengele Keith Siame

Thesis presented in partial fulfilment of the requirements for the degree Mater of Science (Molecular Biology) in the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisor: Prof N.C. Gey van Pittius
Co-supervisor: Prof R.M. Warren
Prof: S.M. Sampson

December 2016

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged.
Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the authorship owner thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature:

Date:

ABSTRACT

Tuberculosis has continued to be a global health concern and warrants an increased understanding of its causative agent *Mycobacterium tuberculosis* (*M. tuberculosis*) in terms of its evolution, virulence and other biological traits. *M. tuberculosis* has been sub-divided into a number of lineage families and sub-lineages based on a number of molecular markers. The Beijing lineage of *M. tuberculosis* has been responsible for a large proportion of tuberculosis cases in Cape Town South Africa. It's evolution and biological characteristics in Cape Town have been investigated using a variety of molecular markers resulting in the identification of 7 sub-lineages. These however have not been investigated as a group using whole genome sequencing. Furthermore, two isolates from sub-lineage 7 reflecting either on-going transmission (clustered) or the absence of transmission (unique) were shown to have a hyper-virulent or hypo-virulent phenotypes in a murine infection model, respectively. Additionally, the hyper-virulent strain elicited an anti-inflammatory TH2 immune response in the murine model whilst the hypo-virulent strain had a pro-inflammatory TH1 immune response. The genetic mechanism(s) underlying these contrasting phenotypes remain to be elucidated.

This study aimed to further elucidate the evolutionary history of the 7 sub-lineages of the Beijing lineage of *M. tuberculosis* in a Cape Town suburb of South Africa using whole genome sequencing analysis.

Whole genome sequencing of the 7 sub-lineages of the Beijing lineage was performed on an Illumina platform generating 105 bp paired-end reads. In addition further sequencing of 2 strains of sub-lineage 7 having contrasting phenotypes was done on a PacBio platform generating long single end raw reads. Three mapping algorithms were used to align the Illumina paired-end reads to the *M. tuberculosis* reference H37Rv. An overlap of SNPs called by each mapping algorithm determined our set of high confidence SNPs which were subsequently used for phylogenetic and comparative SNP analysis. De novo assembly was performed using MIRA and CELERA software to generate a hybrid assembly using Illumina and PacBio raw reads. The super-contig was searched to identify the sequence adjacent to the IS*6110* location. NCBI BLAST was used to determine the location of the IS*6110* element with respect to *M. tuberculosis* H37Rv or *M. bovis* AF2122/97 complete genome reference genomes.

Comparative studies of phylogenetic trees based on genome-wide SNPs showed that the genome-wide SNP tree generated in this study differed from the one based on insertion point markers and selected SNPs in the *mutT* and *ogt* genes by having the evolutionary

positions of sub-lineages 5 and 6 interchanged. The latter markers were however more appropriate for molecular epidemiology studies. The genome-wide phylogenetic trees were also superior to trees based on 43 SNPs in the replication, repair and recombination genes in that the latter exhibited branch collapse in this study.

The comparative SNP analysis among the 7 sub-lineages of Beijing showed the evolution of amino acid changes occurred mostly in the genes of cell wall, cell processes, intermediary metabolism and respiration. Significant overrepresentation of biological processes associated with these changes was however only observed in sub-lineage 1 and observed common ancestor of sub-lineages 1, 2 and 3. Intergenic SNPs unique to each sub-lineage were however identified in close proximity to previously described transcriptional start sites and thus warrant further investigations on their associated transcriptional promoter activity.

The more focused analysis of 2 closely-related members of Beijing sub-lineage 7 having contrasting virulence phenotypes had unique predicted deleterious non-synonymous SNPs which were associated with their whole proteome expression. This included a protein involved in lipid metabolism only expressed in the hyper-virulent strain with the hypo-virulent having a deleterious SNP for the protein and no protein expression. De novo assembly of the two strains also revealed structural variation in the form of a number of unique IS*6110* transposon elements. Of these, 1 IS*6110* element unique to the hypo-virulent strain had an associated large sequence inversion event which has been reported previously by others.

OPSOMMING

Tuberkulose is steeds ‘n wêreldwye gesondheidsprobleem en vereis ‘n beter begrip van die organisme wat dit veroorsaak, *Mycobacterium tuberculosis* (*M. tuberculosis)*, veral met betrekking tot die evolusie, virulensie en ander biologiese eienskappe. *M. tuberculosis* stamme kan op grond van ‘n aantal molukulêre merkers ingedeel word in ‘n aantal linie families en sublinies. Die Beijing linie van *M. tuberculosis* is verantwoordelik vir ‘n groot hoeveelheid van die tuberkulose gevalle in Kaapstad, Suid-Afrika. Die evolusie en biologiese eienskappe van die Beijing linie in Kaapstad is ondersoek deur middel van ‘n verskeidenheid molukulêre merkers, wat gelei het tot die identifikasie van 7 sublinies. Hierdie sublinies is egter nog nie tevore as ‘n groep met behulp van heelgenoom volgordebepaling ondersoek nie. Verder verskil isolate binne sublinies, soos getoon deur twee isolate van sublinie 7 wat onderskeidelik voorturend oordra word (gegroepeer) of nie

oordra word nie (uniek). Daar is ook in ‘n muismodel getoon dat hierdie isolate onderskeidelik hipervirulent en hipovirulent is. Verder het die hipervirulente stam ‘n anti-inflammatoriese TH2 immuunrespons in die muismodel ontlok, waar die hipovirulente stam ‘n pro-inflammatoriese TH1 immuunrespons getoon het. Die genetiese meganisme(s) verantwoordelik vir hierdie kontrasterende fenotipes moet nog verklaar word.

Hierdie studie het verder ten doel gehad om die evolusionêre geskiedenis van die 7 sublinies van die Beijing linie van *M. tuberculosis* in ‘n voorstad van Kaapstad, Suid-Afrika uit te lig deur middel van heelgenoom volgordebepaling. Heelgenoom volgordebepaling van die 7 sublinies van die Beijing linie is op ‘n Illumina platform gedoen wat 105 basispaar gepaarde einde stringe genereer. Verdere volgordebepaling van 2 stamme van sublinie 7 met kontrasterende fenotipes is op ‘n PacBio platform gedoen, wat lang, enkel-stringe genereer. Drie karteringsalgoritmes is gebruik om die Illumina gepaarde einde stringe te pas op die *M. tuberculosis* H37Rv verwysingstam. ‘n Oorvleueling van SNPs wat deur elke karteringsalgoritme aangewys is, het ‘n stel van hoë sekerheid SNPs bepaal, wat vervolgens gebruik is vir filogenetiese- en vergelykende SNP analises. De novo samestelling is met MIRA en CELERA sagteware gedoen om ‘n hibriedsamestelling te genereer van Illumina en PacBio onbewerkte stringe. Ten einde die relatiewe posisies van *IS*6110 ten opsigte van *M. tuberculosis* H37Rv of *M. Bovis* AF2122/97 te bepaal met behulp van NCBI BLAST, is hierdie supersamestelling is areas naasliggend aan *IS*6110, waarvan die volgorde bekend is, geïdentifiseer

Vergelykende studies van filogenetiese bome wat gebaseer is op genoomwye SNPs het getoon dat die genoomwye SNP boom wat in hierdie studie gegenereer is, verskil van die een wat gebaseer is op invoegingspunt merkers en SNPs in die *mutT* en *ogt* gene deurdat die evolusionêre posisies van sublinies 5 en 6 omgeruil is. Laasgenoemde merkers was egter meer toepaslik vir molekulêre epidemiologiese studies. Die genoomwye filogenetiese bome was ook beter as die die bome wat op 43 SNPs in die replikasie, herstel en rekombinasie gene gebaseer is, deurdat laasgenoemde vertakkingsineenstorting veroorsaak het in hierdie studie.

Die vergelykende SNP analise tussen die 7 sublinies van Beijing wys dat die evolusie van aminosuurveranderinge meestal voorkom in die gene van selwand, selprosesse, intermediêre metabolisme en respirasie. Beduidende oorverteenwoordiging van biologiese prosesse wat geassosiëer word met hierdie veranderinge is egter waargeneem in sublinie 1 en die waargenome gemene voorsaat van sublinies 1, 2 en 3. Intergeniese SNPs wat uniek is tot elke sublinie is egter waargeneem in posisies wat naby geleë is aan voorheen beskryfde transkripsiebeginpunte, en oorloof verdere navorsing oor die

geassosiëerde transkripsionele promoter aktiwiteit.

Die meer gefokusde analise van 2 nabyverwante lede van Beijing sublinie 7 wat kontrasterende virulensie fenotipes het, het verskillende voorspelde nadelige nie-sinonieme SNPs getoon, wat verband hou met hul heel-proteoom uitdrukking. Dit sluit ‘n lipiedprotein in wat net in die hipervirulente stam uitgedruk word, terwyl die hipovirulente stam ‘n nadelige SNP vir die proteïen gehad het, met geen proteïenuitdrukking nie. De novo samestelling van die twee stamme het ook strukturele variasie in die vorm van ‘n aantal unieke IS*6110* transposon elemente onthul. Een van hierdie IS*6110* elemente wat uniek was tot die hipovirulente stam, het ‘n geassosiëerde groot volgorde inversie gehad wat voorheen deur ander outeurs beskryf is.

LIST OF ABBREVIATIONS

°C	Degrees Celcius
AA	Amino acid
AIDS	Acquired immune deficiency syndrome
bam	binary alignment map
BCG	Bacillus Calmette-Guérin
BFAST	Blat-like fast accurate Search tool
BLAST	Basic local alignment search tool
bp	Base pair
BWA	Burrows-Wheeler Aligner
CAS	Central Asian strains
CDS	Coding sequence
DNA	Deoxyribo-nucleic acid
DNase	Deoxyribonuclease
DNS	Deoksieribo-nukleïensuur
dNTP	Deoxynucleoside triphosphate
DR	Drug resistant, direct repeat
DS	Drug sensitive
DTT	Dithiothreitol
DVR	Direct variable repeat
EAI	East African Indian
et al.	*et al* (and others)
GATK	Genome analysis toolkit
HIV	Human immunodeficiency virus
i.e.	id est (that is)
IL	Interleukin
In/del	Small insertions and deletions
IS*6110*	Insertion sequence *6110*
LAM	Latin-American Mediterranean
LC	Liquid chromatography
LCC	Low copy clade
LSP	Large sequence polymorphism
M.	*Mycobacterium*

MDR	Multi drug resistant
MIRU/VNTR	Mycobacterial interspersed repeat elements / variable number tandem repeats
ml	Milliliter
mm	Milimeter
mM	Millimolar
MS	Mass spectrometry
MTBC	*Mycobacterium tuberculosis* complex (excluding *M. canettii*)
MTC	*Mycobacterium tuberculosis* complex (including *M. canettii*)
MW	Molecular weight
NHLS	National Health Laboratory Service
NGS	Next generation sequencing
nt	Nucleotide
OD	Optical density
ORF	Open reading frame
PCR	Polymerase chain reaction
PGG	Principle genetic group
pH	Potential of hydrogen
QS	Quality score
RD	Region of difference
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic-acid
sam	sequence alignment map
SAWC	South Africa Western Cape
SB	Sodium borate (buffer)
SNP	Single nucleotide polymorphism
Spoligotyping	Spacer oligonucleotide typing
TB	Tuberculosis
TCA	Tri-carboxylic acid
TDR	Totally drug resistant
T_m	Melting temperature
TNF	Tumour necrosis factor
vcf	Variant call format

WGS	Whole genome sequencing
WHO	World Health Organization
www	Wold wide web

Table of Contents

1 LITERATURE REVIEW

The human pathogen *Mycobacterium tuberculosis* (*M. tuberculosis*) is a member of the *Mycobacterium tuberculosis* Complex (*MTBC*) which consists of *M tuberculosis*, *Mycobacterium africanum, Mycobacterium bovis, Mycobacterium canetti, Mycobacterium microti, Mycobacterium pinnipedii,* dassie bacillus, *Mycobacterium* orygis, chimpanzee bacillus and *Mycobacterium caprae*. The *MTBC* members have evolved from the same common progenitor (Gutierrez *et al*., 2005; Helal *et al*., 2009; Wirth *et al*., 2008) and share 99% similarity at the genome level. Published works have estimated that speciation occurred between 20,000-35,000 years ago from a common ancestor termed *M. prototuberculosis* (Gutierrez *et al*., 2005; Wirth *et al*., 2008). The use of whole genome sequencing on a global selection of strains now estimates speciation to have occurred about 70,000 years ago (Comas *et al*., 2013). An out of Africa migration of early hominids into Asia and subsequently to the rest of the world allowed for the evolution of distinct lineages of *M. tuberculosis* within defined populations. Population mixing occurred more recently, especially with the development of land and sea trade routes (Comas *et al*., 2013; Hershberg *et al*., 2008; Wirth *et al*., 2008). The above mentioned scenarios and present day travel have shaped the current distribution of *M. tuberculosis* strains. The epidemiological significance of different strain lineages in different geographical areas is thought to reflect host-pathogen compatibility (Gagneux *et al*., 2006; Reed *et al*., 2009) .

This review describes the genetic, transmission and geographic description of *M. tuberculosis* lineages predominantly responsible for tuberculosis in Eastern and Southern Africa as well as the Indian sub-continent and East Asia.

1.1 Genetic markers and description of lineages

A number of genotyping methods have been used in epidemiological, phylogenetic and evolutionary studies to investigate strain variation and distribution. These methods are based on genomic changes which can either be single base changes involving synonymous or non-synonymous single nucleotide polymorphisms (SNPs or nSNPs), large sequence changes involving insertions, duplications and deletions and expansion and contraction of repetitive regions (Alland *et al*., 2007; Filliol *et al*., 2006; Sreevatsan *et al*., 1997). Each method has its own merits and disadvantages and a combination of methods are often used in studies in order to increase the discriminatory index (Gutierrez *et al*., 2006; Mathuria *et al*., 2008; Narayanan *et al*., 2008; Sola *et al*., 2003).

The molecular typing techniques that have been used to describe the different strains of the *M. tuberculosis* include Restriction Fragment Length Polymorphism (RFLP) (Warren *et*

al., 1999), spoligotyping (Kamerbeek *et al*., 1997) region of difference (RD) analysis (Gagneux *et al*., 2006; Tsolaki *et al*., 2004), variable number tandem repeat (VNTR) typing (Supply *et al*., 2006) and single nucleotide polymorphism (SNP) analysis (Gutacker *et al*., 2006). Restriction Fragment Length Polymorphism (RFLP) based on the location and number of the transposon element IS*6110* is the gold standard for molecular epidemiology of tuberculosis (Jagielski *et al*., 2014; van Soolingen and Kremer, 2009). Strain lineages have a characteristic IS*6110* RFLP pattern, which can aid the identification of a strain which has acquired a favorable transmission phenotype in a specific geographical area (van der Spuy *et al*., 2009; Warren *et al*., 1999). Spoligotyping is based on variation at the variable Direct Repeat locus of *M. tuberculosis*. The presence and absence of 43 unique spacer sequences among direct repeats is associated with specific strain lineages (Streicher *et al*., 2007). Region of difference (RD) analysis is based on the deletion of large sequences. The RDs may be unique to a specific strain lineage/sub-lineage or present among different strain families suggesting their relatedness. Those corresponding to unique events are of particular importance in evolution and phylogenetic studies (Alland *et al*., 2007). The currently recommended VNTR analysis utilizes a set of 24 loci (Supply *et al*., 2006). Sub-sets of this have been demonstrated to provide epidemiological and phylogenetic information unique to strain families (Brown *et al*., 2010). The decrease in cost of Next Generation Sequencing in recent years has provided an opportunity to use the entire genome for analysis of strains. Bioinformatics has subsequently become an important tool in analysing the increasing amount of data that is being generated by high throughput sequencing techniques (Bravo and Procop, 2009; Quail *et al*., 2012; Roetzer *et al*., 2013).

1.2 Genomic description of Mycobacterium tuberculosis lineages

Polymorphisms at the *katG* 95 and *gyrA* 463 positions of the *M. tuberculosis* genome has been used to group strains into 3 principle genetic groups (PGGs) (Sreevatsan *et al*., 1997). PGG1 is evolutionarily the oldest followed by PGG2 and 3, respectively. PGG1 members are responsible for most of the tuberculosis related cases in Asia, particularly on the Indian sub-continent and in the Far East. Also affected by this group of strains is the eastern coast of Africa including Madagascar, as well as the southern parts of the continent. More than 50% of the global TB case load has been attributed to this group of strains making them an important group to study in relation to the control and eradication of TB. The PGG1 strains have further been sub-divided into different lineages using the above mentioned molecular typing techniques of RFLP, spoligotyping, RD analysis and

SNP typing. The group is split into 4 spoligotype lineages, namely; Manu, East African Indian (EAI), Central Asian (CAS) and Beijing (Brudey *et al*., 2006; Comas *et al*., 2010; Filliol *et al*., 2006; Flores *et al*., 2007; Gagneux *et al*., 2006; Reed *et al*., 2009). To our knowledge no study has investigated the characteristics of PGG1 members to compare the genotypes and to investigate how these translate to observed phenotypes. Deciphering how these traits evolved across the PGG1 members would increase our understanding of the biology this pathogen. Studies have primarily focused on only one lineage within the group, namely the Beijing lineage. This lineage has emerged as a cosmopolitan strain lineage causing disease in humans across the globe, hence the more abundant literature on this lineage compared to the other PGG1 lineages. However, members of the other PGG1 lineages also contribute to the TB disease burden, particularly in the developing world, and these warrant more attention in terms of research as the effort to control and eradicate TB continues.

1.3 Spoligotype Analysis

Manu Lineage

The Manu lineage was initially described in India following the discovery of a strain harbouring a near full set of intact spacer sequences, except spacers sequences at positions 33 and 34, as shown in Figure 1 (Singh *et al*., 2004). Following the discovery of the Manu strain with the above mentioned spoligotype pattern, strains with spacer 34 deleted only as well as spacers 34, 35 and 36 deleted were discovered (Brudey *et al*., 2006). These were named Manu 1 and 3 respectively and strains with only spacers 33 and 34 subsequently named Manu 2 as depicted in Figure 1.1-a. Variants of the 3 have been reported in several areas, with the greatest numbers in Egypt, Ethiopia and India (Belay *et al*., 2014; Brudey *et al*., 2006; Chatterjee *et al*., 2010; Helal *et al*., 2009; Thomas *et al*., 2011).

A number of reports have however cautioned the interpretation of a Manu spoligotype, as this hybridization pattern can be created from the superimposition (mixed infection) of two strains e.g. Beijing and T lineage (see Figure 1.1-b) (Viegas *et al*., 2010). This can only be verified when RD analysis and spoligotyping are both used (Viegas *et al*., 2010). However, few of Manu strains reported in the SITVIT database have both spoligotyping and RD analysis done (Demay *et al*., 2012; Flores *et al*., 2007; Thomas *et al*., 2011). Thus, it should be questioned as to whether these strains are true Manu spoligotypes.

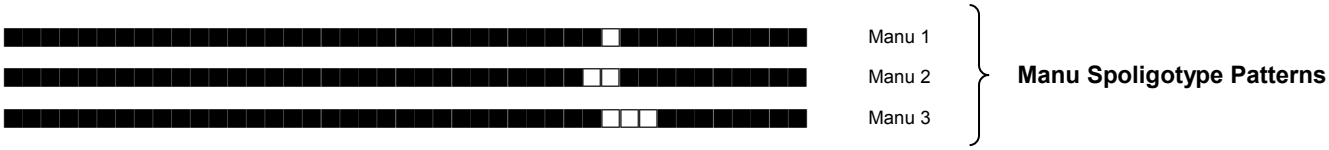


Figure 1.1-a: Manu spoligotype patterns as described by Brudey *et al* 2010.

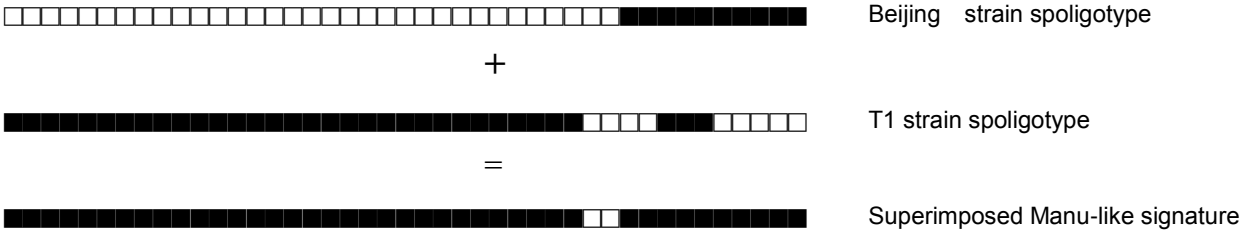


Figure 1.1-b: Possible scenario of how a mixture of two strains can reveal a "Manu2"-like spoligotype signature.

EAI Lineage

The EAI spoligotype signature has the DR spacers 29-32 and 34 deleted. Strains harbouring this signature can be grouped into 9 sub-classes (see Figure 1.2) (Brudey *et al*., 2006). Some of the EAI spoligotypes have been named after the geographical location where they were isolated.

Family	Shared Type No.	Spoligotype Pattern
EAI-5	236	
EAI 1-SOM	48	
EAI2-Manilla	19	
EAI2-Nonthaburi	89	
EAI3-IND	11	
EAI4-VNM	139	
EAI6-BGD	1591	
EAI6-BGD	11898	
EAI8-MDG	109	

Figure 1.2: EAI spoligotypes (Brudey *et al*., 2006; Phyu *et al*., 2009).

CAS Lineage

The Central Asian (CAS) lineage is characterised by the absence of DR spacers 4-7 and 23-34 (Brudey *et al*., 2006, Demay *et al* 2012). Four sub-lineages are described in the SpolDB4 and SITVIT databases with deletions as indicated in Figure 1.3.

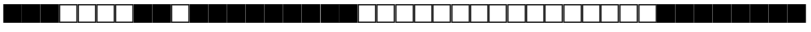
Strain	Shared Type No.	Spoligotype Pattern
CAS1-Delhi	26	
CAS1-Kili	21	
CAS1-variant	25	
CAS2	288	

Figure 1.3: CAS Spoligotypes.

Beijing Lineage

The classic spoligotype pattern of the Beijing lineage is characterised by the deletion of spacers 1-34 with only the last 9 spacers being present. Variations of Beijing exist where additional spacers are deleted from the classic 9 spacers of Beijing family spoligotypes (Brudey *et al*., 2006). More recently, an ancestral form of the Beijing lineage was described where all 43 spacers are present as is illustrated in Figure 1.4 (Flores *et al*., 2007).


Classical Beijing


Ancestral Beijing

Figure 1.4: Beijing classical spoligotype pattern with spacers 1 to 34 deleted and ancestral Beijing with all 43 spacers intact.

1.4 RD Analysis

The earliest sub-grouping of PGG1 members by RD analysis was by the TbD1 deletion (Brosch *et al*., 2002). This deletion divides PGG1 into two groups with members retaining the region termed “ancestral or ancient” strains and those having the deletion termed “modern” strains (Brosch *et al*., 2002). PGG1 is the only group that has TbD1^{+} members among *M. tuberculosis* strains. Both Manu and EAI strains are classified as ancestral strains due to the presence of the TbD1 sequence, while CAS and Beijing strains are classified as modern strains due to the absence of the TbD1 sequence. These 4

spoligotype lineages can further be grouped by lineage-specific RDs (Figure 1.5) (Flores *et al*., 2007; Gagneux *et al*., 2006; Reed *et al*., 2009; Thomas *et al*., 2011; Tsolaki *et al*., 2004). The Manu and EAI (lineage 1) have the RD239 deletion and are collectively termed the Indo-Oceanic lineage. CAS is defined as the Indian East Africa lineage according to the RD750 deletion whilst strains having the RD105 deletion are classified as the East Asian lineage. The Beijing spoligotype lineage described earlier is part of the East Asian lineage which has additional members not having the classic Beijing lineage spoligotype pattern. This includes isolates having the full set of spoligotype spacers (Flores *et al*., 2007) as well as strains having additional deletions in the remaining 9 spacers of the classic Beijing spoligotype pattern (Brudey *et al*., 2006). RD-based lineages can be further resolved into sub-lineages based on additional large deletions for the Beijing lineage (Schürch *et al*., 2011a; Tsolaki *et al*., 2004, 2005). These include RD142, RD150, RD181 and RD207.

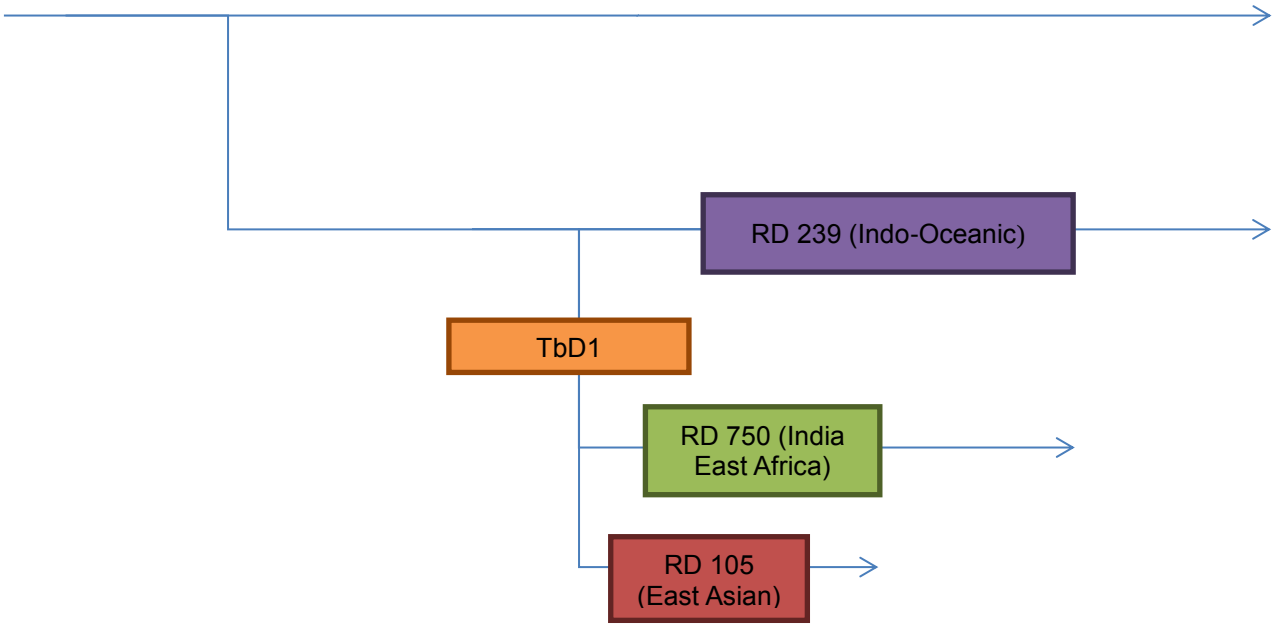


Figure 1.5: Diagram illustrating lineage defining deletions in East Asian, India East Africa and Indo Oceanic lineages.

Some RDs are deleted multiple times/independently in strains harbouring different lineage specific RDs and are thus not necessarily phylogenetically relevant, and may contribute greatly to genetic diversity within the MTBC (Tsolaki *et al*., 2004). These included RD149 and RD152 (Kanji *et al*., 2011a; Tsolaki *et al*., 2004).

1.5 SNP Analysis

SNPs have been used to differentiate *M. tuberculosis* strains and other members of the MTBC as shown in Figure 1.6. This is currently considered the gold standard to identify

evolutionary events and to infer phylogenies (Abadia *et al*., 2010; Baker *et al*., 2004; Chuang *et al*., 2010a; Comas *et al*., 2010; Hershberg *et al*., 2008; Schürch *et al*., 2011b; Sreevatsan *et al*., 1997). Within PGG1, lineage-specific SNPs have been identified in individual genes involved in replication, repair and recombination (3R) and in cell wall biosynthesis-associated genes as shown in Table 1.1. (Abadia *et al*., 2010; Chuang *et al*., 2010a, 2010b; Mestre *et al*., 2011).

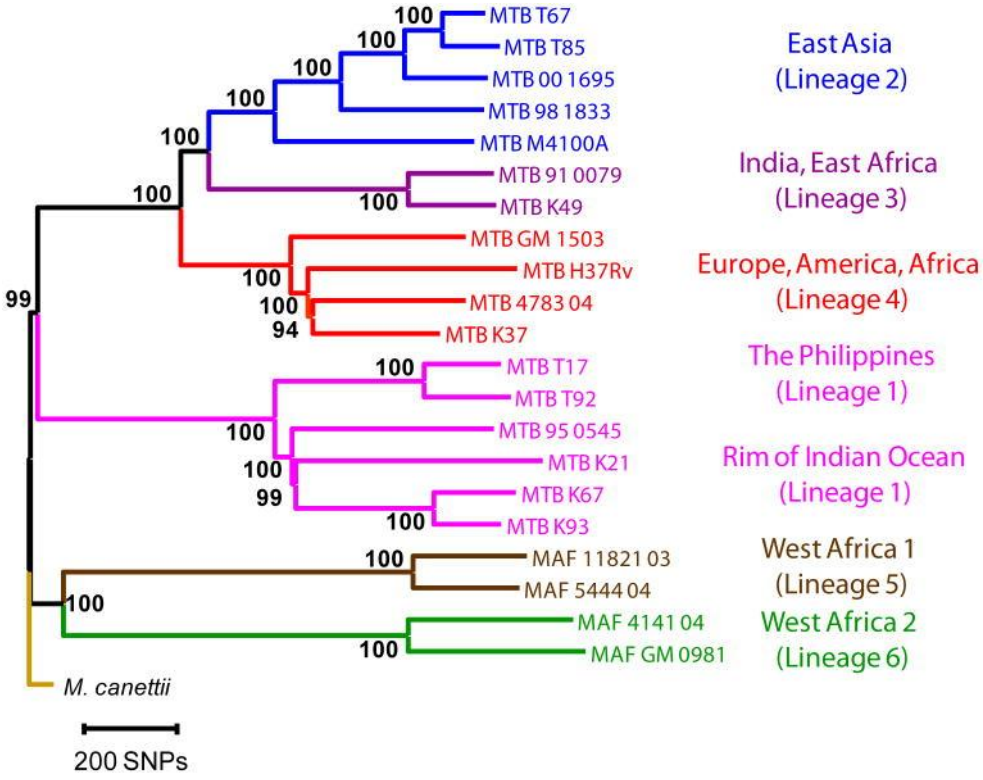


Figure 1.6: Neighbour-joining phylogeny based on 9,037 variable common nucleotide positions across 21 human *M. tuberculosis* complex genome sequences. The tree is rooted with *M. canettii*, the closest known outgroup. Node support following 1,000 bootstrap replications is indicated. Branches are coloured according to the six main phylogeographic lineages of MTBC. Highly congruent topologies were obtained by Maximum likelihood and Bayesian inference (*Comas et al., 2010)*.

Table 1.1: PGG1-defining SNPs in individual genes with the exception of the Manu lineage.

Clade	Clade Specific Mutations Showing Gene and Gene Position of SNP						
	***PimB*^{270}**	***FbpA*^{156}**	***FbpA*^{4}**	***FbpB*^{238}**	***FbpC*^{158}**	***fad28*^{507}**	***recO*^{606}**
Indo-Oceanic lineage 1	GCC→GTT						GGC→GGT
EAI lineage 3			GTT→GTC		GGC→AGC		
East Asian Lineage 2 (Beijing Ancestral)		AGG→ATG					
East Asian Lineage 2 (Beijing Modern)				CCC→CCA			
East Asian Lineage 2 (All Beijing)						ATC→ATT	

The advantage of analysing SNPs is that this information is sufficiently robust to enable the accurate grouping of strains independently of their spoligotype signatures (i.e. identifying Beijing genotype isolates having all 43 DR spacers present) (Chuang *et al*., 2010b). SNP based analysis has also been performed utilising numerous SNPs from a large number of strains and representative of the global *M. tuberculosis* landscape (Comas *et al*., 2010). With the increase in the number of genomes that have been sequenced, studies have incorporated a greater number of SNPs in elucidating the phylogeny and evolution of MTBC members. Phylogenetic scenarios inferred by genome wide SNPs and use of 89 genes in the replication, repair and recombination system 3R SNPs largely agree (Comas *et al*., 2009, 2010; Hershberg *et al*., 2008). Degenerate sets of lineage-defining SNPs (see Table 1.2) have been identified and the resulting phylogenetic tree predicted that the Indo-Ocenaic lineage 1 consisted of two sub-lineages: “The Philipines” and “Rim of the Indian Ocean” sub-lineages. In agreement with phylogenenomic studies by others, The Indo-Oceanic lineage 1 is the most ancestral of PGG1 members followed the India East Africa lineage 3 and the East Asian lineage 2, respectively (Filliol *et al*., 2006; Gutacker *et al*., 2006). In summary, SNP analysis shows that the PGG1 grouping of *M. tuberculosis* members is much more diverse than previously thought despite the high clonal nature of *M. tuberculosis*.

Table 1.2: Lineage Specific SNPs of PGG1 and Associated Region of Difference (RD) Marker.

Lineage/ Sublineage	SNP Name	*Rv Number*	Nucleotide Position	H37Rv Nucleotide	Mutant Nucleotide	Codon	H37Rv Amino Acid	Mutant Amino Acid	Lineage/ sublineage defining genomic deletion
Lineage 1	Rv0005_0990n	*0005*	0990	atG	atC	0330	Met	Ile	RD239
	Rv0006_1151n	*0006*	1151	gCa	gTa	0384	Ala	Val	
	Rv0410c_1842s	*0410c*	1842	tcG	tcA	0614	Ser	Ser	
	Rv0934_1022n	*0934*	1022	aCc	aTc	0341	Thr	Ile	
	Rv1996_0052n	*1996*	0052	Ccc	Tcc	0018	Pro	Ser	
	Rv2462c_1086s	*2462c*	1086	gaT	gaC	0362	Asp	Asp	
	Rv3132c_1680s	*3132c*	1680	gcG	gcC	0560	Ala	Ala	
	Rv3221c_0085n	*3221c*	0085	Gtc	Atc	0029	Val	Ile	
"Manila"	Rv0006_1959s	*0006*	1959	ctG	ctC	0653	Leu	Leu	none
	Rv0164_0415n	*0164*	0415	Ctg	Atg	0139	Leu	Met	
	Rv0288_0028n	*0288*	0028	Gcg	Acg	0010	Ala	Thr	
	Rv0410c_2117n	*0410c*	2117	tTc	tCc	0706	Phe	Ser	
	Rv1009_0724n	*1009*	0724	Gag	Aag	0242	Glu	Lys	
	Rv1996_0157n	*1996*	0157	Ggg	Cgg	0053	Gly	Arg	
	Rv2030c_1137s	*2030c*	1137	ctG	ctA	0379	Leu	Leu	
	Rv2031c_0426s	*2031c*	0426	tcC	tcT	0142	Ser	Ser	
	Rv3261_0015s	*3261*	0015	gtT	gtC	0005	Val	Val	
Lineage 3	Rv0129c_0472n	*0129c*	0472	Ggc	Agc	0158	Gly	Ser	RD750
	Rv2959c_0219n	*2959c*	0219	gaG	gaT	0073	Glu	Asp	
	Rv3133c_0419n	*3133c*	0419	gCc	gGc	0140	Ala	Gly	
	Rv3804c_0012s	*3804c*	0012	gtT	gtC	0004	Val	Val	
Lineage 2	Rv2952_0526n	*2952*	0526	Ggg	Agg	0176	Gly	Arg	RD105
"Non-Beijing "	Rv0652_0328n	*0652*	0328	Aag	Gag	0110	Lys	Glu	none
	Rv1173_1513n	*1173*	1513	Gcg	Acg	0505	Ala	Thr	
	Rv1996_0153s	*1996*	0153	ccG	ccT	0051	Pro	Pro	
	Rv2330c_0487n	*2330c*	0487	Gat	Aat	0163	Asp	Asn	
	Rv2957_0411s	*2957*	0411	ctG	ctA	0137	Leu	Leu	
"Beijing"	Rv2450c_0059n	*2450c*	0059	aCg	aGg	0020	Thr	Arg	RD207

1.5.1 *Increased Resolution of the Beijing lineage*

The lineage 2 East Asian Beijing lineage remains the most studied strain family with respect to genetic markers. Different studies have used varying permutations of genetic markers to delineate sub-lineages of the Beijing lineage using either strains from different geographical areas or strains circulating in a defined region, (Dos Vultos *et al*.; Hanekom *et al*., 2007a; Mestre *et al*., 2011; Schürch *et al*., 2011b; Merker *et al.,* 2015; by Luo T *et al*.,2015). Typical (modern) Beijing strains can be distinguished from Atypical (ancient or ancestral) Beijing strains by the presence of an IS*6110* element in the NTF region which is depicted in Figure 1.7 and whose primers are shown in Table 1.3 (Plikaytis *et al*., 1994; Schürch *et al* 2011).

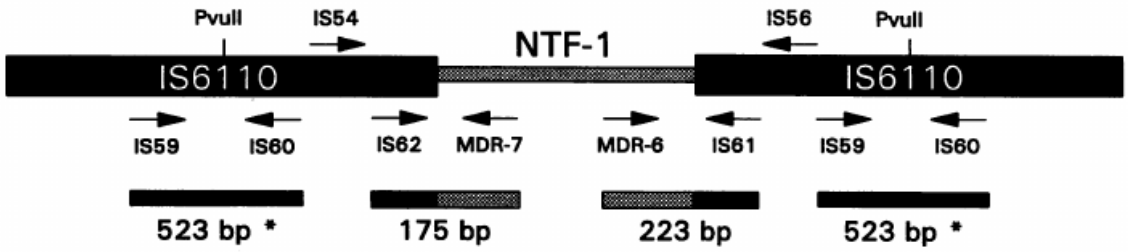


Figure 1.7: Schematic representation of IS*6110* direct repeat with the intervening NTF-1 sequence, the location of the oligonucleotides used in the PCR, and the expected products from strain W. *, positive internal PCR control product. (Plikaytis *et al*., 1994).

Table 1.3: Sequences of oligonucleotides used as primers for NTF region (Plikaytis *et al*., 1994).

Primer	Target	Sequence (5'-3')
IS54	IS*6110*	TCGACTGGTTCAACCATCGCCG
IS56	IS*6110*	GCGACCTCACTGATCGCTGC
IS59	IS*6110*	GCGCCAGGCGCAGGTCGATGC
IS60	IS*6110*	GATCAGCGATCGTGGTCCTGC
IS61	IS*6110*	GACCGCGGATCTCTGCGACC
IS62	IS*6110*	ACCAGTACTGCGGCGACGTC
MDR-6	NTF-1	CCAGATATCGGGTGTGTCGAC
MDR-7	NTF-1	CGCGAGATCTCATCGACAACC

Further resolution of the Beijing lineage has been demonstrated by using an array of different markers as described below. Beijing family members can be resolved into 5 sub-lineages based on RD analysis (Gagneux *et al*., 2006; Tsolaki *et al*., 2005). All members have the RD105 deletion with subsets of these having further deletions as exemplified in Figure 1.8.

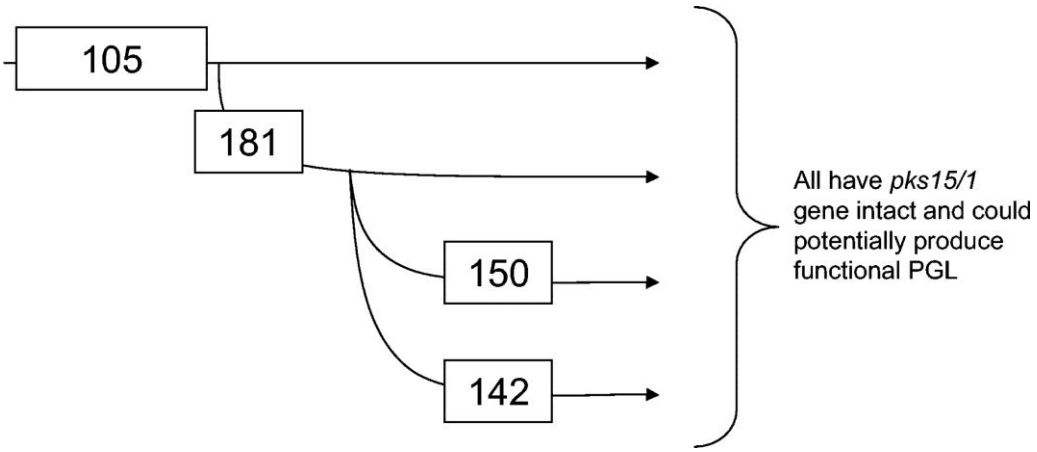


Figure1.8: Beijing/W family of *M. tuberculosis* is monophyletic (Tsolaki *et al*., 2005).

Variation in the DNA repair genes *mutT2*, *mutT4,* and *ogt* has been observed for the Beijing lineage (Ebrahimi-Rad *et al*., 2003). Additionally, an investigation of SNPs in the replication repair and recombination (3R) system genes of *M. tuberculosis* from a global representation of *M. tuberculosis* strains revealed that 22 genes were polymorphic for Beijing strains (Comas *et al*., 2009). Sequencing of these 3R genes from a global set of Beijing and other lineage family strains identified 41 SNPs that were specific for the Beijing lineage. Of these, 30 SNPs enabled the discrimination of 24 sequence types among Beijing strains. Amongst the sequence types identified, a group or node referred to as Bmyc10 was found to be the most prevalent in a larger set of global Beijing isolates followed by a group identified as Bmyc25. In addition to forming the largest cluster, Bmyc10 had a large global distribution when compared to other identified groupings (Dos Vultos *et al*.; Mestre *et al*., 2011).

An analysis of South African isolates using previously described SNPs, RDs and insertion sites for IS*6110* was undertaken by (Hanekom et al., 2007a) to describe the evolution of Beijing strains (Hanekom et al., 2007a). Seven sub-lineages were identified from the resulting phylogenetic tree. Interestingly, that study showed that isolates from the different sub-lineages transmitted at different rates. The topology of the phylogenetic tree was supported by a whole genome comparative analysis (Schürch *et al*., 2011).

A number of sub-divisions of the Beijing lineage have been identified following WGS. These include the study by Merker *et al*., (2015) where 3 ancestral (Asia Ancestral 1-3) and five modern groupings (Central Asian, European-Russian, Pacific, Asian Africa 1 and 2) were identified from 7 Clonal Complexes based on MIRU-VNTR typing. Three monophyletic groups were identified on the other hand in the study by (Luo *et al*., (2015)

and 27 clonal complexes by (Schürch et al., 2011a).

1.6 Immunogenicity of PGG1 Members

The human immune response to *M. tuberculosis* infection is initiated by its uptake by macrophage immune cells. Following the uptake of *M. tuberculosis* by macrophages, during the innate immune response, antigens of *M. tuberculosis* are presented to T-cells in the adaptive phase of the immune response involving an interplay of T-helper (T_H) cells as illustrated in Figure 1.9. To date, no study has been done to collectively characterize the immune responses elicited during infection with the different PGG1 members. Most studies report the analysis of one or two members and are largely restricted to early time points following infection. Only limited data is available for members of the Manu lineage.

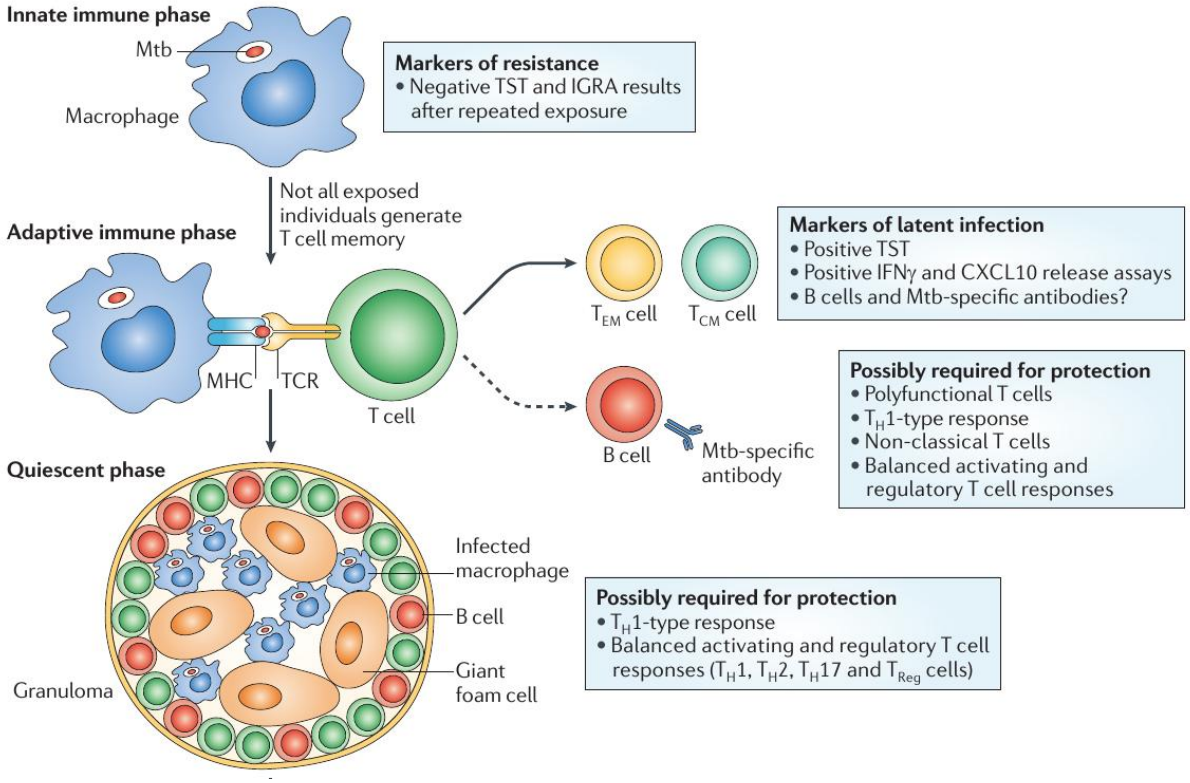


Figure 1.9: The stage of infection is determined by the ability of the host innate and adaptive immune systems to eradicate or control *M. tuberculosis*. In the adaptive immune phase, T cells are engaged by antigen-presenting cells, and this generates effector and memory T cells (both effector memory T (TEM) and central memory T (TCM) cells). An optimal T helper (TH) cell balance is required to control *M. tuberculosis* while limiting immunopathology. This balanced reaction includes pro-inflammatory TH1-type responses (characterized by interferon-γ (IFNγ), tumour necrosis factor (TNF) and interleukin-12 (IL-12) production) and TH17-type responses (characterized by IL-17 production). However, it also involves TH2-type responses (associated with IL-4 production) and regulatory T (TReg) cell phenotypes that limit immunopathology. *(Adapted from Walzl et al 2011)*.

Members of the Indo-Oceanic lineage (Manu and EAI) demonstrate varied immunological responses differing from the Beijing lineage in some studies in that they generally exhibit a pro-inflammatory cytokine response (Homolka *et al*., 2010; Portevin *et al*., 2011; Rakotosamimanana *et al*., 2010; Wang *et al*., 2010). Cytokines TNFα and IL-10 have been previously shown to play a role on the control of *M. tuberculosis* infection (Beamer *et al*., 2008; Cavalcanti *et al*., 2012; Dambuza *et al*., 2008; Redford *et al*., 2011). However, an anti-inflammatory signature similar to that observed for the Beijing clade was also detected for some EAI strains suggesting heterogeneity in the immunological responses of the EAI clade (Homolka *et al*., 2010; Portevin *et al*., 2011; Rakotosamimanana *et al*., 2010; Wang *et al*., 2010).

Similarly, heterogeneous immunological responses have been reported for members of the CAS clade (Portevin *et al*., 2011). The RD750 deletion of the CAS clade, has been linked to the induction of a non-protective anti-inflammatory immune response in a UK study (Newton *et al*., 2006). However, CAS strains may also harbour the RD149 deletion or concurrent RD149 and RD152 deletions(Kanji *et al*., 2011a, 2011b). The latter strains were shown to elicit more TNFα than those strains having only the RD750 deletion. The physiological effect of these additional deletions was hypothesised to be linked to the uptake of the pathogen at early time points as there was an associated up-regulation of IL-10 (Kanji et al., 2011a). Interestingly, a sub-set of Beijing strains have also been shown to harbour RD149 and RD152 deletions (Kanji et al., 2011a). However, the immunological impact of these deletions within this genetic background remains to be elucidated.

A comparative analysis of cytokine and chemokine molecules induced in macrophages and dendritic cells showed that there was a homogeneous response with respect to the different sub-lineages of Beijing (defined according to polymorphisms in the 3R genes) (Wang *et al*., 2010). The immune response elicited in these experiments did not favour a protective pro-inflammatory response. However, the authors stated that the results could be different at later time points during infection and in an *in vivo* model. An example of the latter was demonstrated in a mouse infection model experiment where the transcriptome was found to be heterogeneous for anti- and pro-inflammatory cytokines for strains belonging to different Beijing sub-lineages. Furthermore, the strain lineages that elicited an anti-inflammatory response also exhibited high virulence by the killing of mice (Aguilar *et al*., 2010). What was even more striking however was that 2 strains from the same sub-lineage exhibited hypo- and hyper-virulence. Homolka *et al*., 2010 also showed that inherent immune responses elicited by strains are linked to genomic background. This was

more evident when the investigations were done in activated macrophages. The results of Homolka and Aguila taken together thus point to a differential expression of pro-inflammatory cytokines induced by Beijing strains when done in either an *in vivo* model or activated macrophages. This, in turn suggests that the immunological responses induced by members of the Beijing sub-lineages are more diverse than reported by Wang *et al*., 2010.

1.7 Host-Pathogen association in PGG1 members

The global spread of different lineages of *M. tuberculosis* have exhibited a phylo-geographical pattern and some evidence has pointed to an association of specific strains with populations originating in different parts of the world (Gagneux, 2012; Reed *et al*., 2009). PGG1 member strains are more prevalent in Asia and the coastal areas of the Eastern and Southern parts of Africa (Brudey et al., 2006; Ferdinand et al., 2005; Hanekom et al., 2007a; Kibiki et al., 2007; van der Spuy et al., 2009; Viegas et al., 2010; Warren et al., 2004). It has been hypothesised that the introduction of PGG1 strains along the east coast of Africa resulted from sea trade between Asia and Africa (Brudey *et al*., 2006; Hershberg *et al*., 2008; Schürch *et al*., 2011b; Wirth *et al*., 2008). The origin of the Beijing lineage in this regard has been shown to originate from China and spread to other regions of the world (Luo *et al*., 2015; Merker *et al*., 2015). This would have seeded the strains now prevalent in East and Southern Africa with little evidence suggesting that they were more likely to infect only those people where the strains originated from than the indigenous populations. It is worth pointing out that studies like the ones done in Montreal and San Francisco (Gagneux *et al*., 2006; Reed *et al*., 2009) have not been done in these TB endemic areas. However, the efficient spread of Beijing sub-lineage 7 strain in the South African Coloured population in Cape Town, South Africa suggests host-pathogen compatibility and was supported by the association between certain HLA types and Beijing sub-lineage 7 strains (Salie *et al*., 2014). A founder effect to the spread of PGG1 members in Africa rather than an association with people originating in the same areas as the strains can also not be ruled out. To this end, the distinct distribution of CAS and EAI in India as well as Madagascar can have its merits in the founder population effect (Arora *et al*., 2009; Ferdinand *et al*., 2005; Singh *et al*., 2004, 2007).

2 MATERIALS AND METHODS

2.1 Overview

The materials and methods chapter for this study gives an outline of the molecular methods and bioinformatics analyses done as part of the current study. An overview of the next-generation sequencing techniques and bioinformatics terminology with respect to next generation whole genome sequence data analysis is also given, where relevant.

The work described in this thesis was done as part of a large on-going project which received ethical approval from the Stellenbosch University Health Research Ethics Committee under the title: An investigation into the evolutionary history and biological characteristics of the members of genus *Mycobacterium*, with specific focus on the different strains of *M. tuberculosis*, other members of the *M. tuberculosis* complex and non-tuberculosis Mycobacteria (NTM), ethics reference number: N10/04/126.

The whole genome sequencing data analysis pipeline described here was developed to analyse the data generated for this project, as well as other ongoing research projects at the time. All bioinformatics was done by the candidate unless stated otherwise (e.g. scripts written by colleagues). Scripts written for the purpose of the analyses done in this study are included in Supplemental data.

2.2 Molecular Methods

2.2.1 Sample Collection

For this study, samples were available from a longitudinal database. Isolates in the database were collected during the period from 1996-2008 from patients resident in Cape Town, Western Cape Province of South Africa. Isolates representing different lineages of *M. tuberculosis* were selected according to their spoligotyping and IS*6110* RFLP patterns (van Embden *et al*., 1993; Kamerbeek *et al*., 1997). Furthermore, representative members of the 7 sub-lineages of Beijing were selected, including 2 members of sub-lineage 7 which had shown contrasting phenotypes in previous studies (Aguilar *et al*., 2010).

2.2.2 Spoligotyping

Spoligotyping was done according to the international standardized protocol (Kamerbeek *et al*., 1997) and assigned to types and families according to SpolDB4 (Brudey *et al*., 2006).

2.2.3 Region of Difference (RD) Analyses

Genome positions for lineage-specific regions of difference (RDs) were obtained from

literature reviewed (Gagneux *et al*., 2006; Tsolaki *et al*., 2005). Bioinformatics and primer design software were used to design specific primers for the detection of the presence or absence of the RDs. The regions of difference were identified in the genome sequence data obtained from the TubercuList web site (http://genolist.pasteur.fr/TubercuList/genome.cgi) relative to the reference strain H37Rv. Primers were designed to detect both the presence and absence of the deletion region. The size of the products generated for deletion and non-deletion sequences were designed to be of different sizes for ease of detection by agarose gel electrophoresis. The primer sets used are given in Table 2.1.

2.2.4 PCR Conditions for RD Analysis

The following conditions were used for PCR amplification for the RD analysis. For a 1x 25 µL reaction, reagents used were as follows: 12.37 µL water, 2.5 µL 10x Buffer, 2 µL MgCl2, 1 µL dNTPs, 0.125 µL Hot Star Taq (Qiagen) and 5 µL Q-solution and a 1.5 µL primer mix of forward, reverse and reverse internal primers. The PCR was done at 95ºC for 15 min where after 35 cycles of 94^{o}C for 35 sec, 60^{o}C for 35sec, 72^{o}C for 55sec, 72^{o}C for 10min, 4^{o}C for ∞. The following lineage specific primers were used on our samples: RD 239 (specific for Manu and EAI), RD 750 (CAS specific), RD 105 (Beijing specific), and Tuberculosis Specific Deletion 1 (TbD1) (separates modern from ancestral strains).The products were run on a 0.8% agarose gel in buffer stained with ethidium bromide. Five microlitres of PCR product was mixed with 5 µL of loading buffer and 5 µL of this mixture loaded onto the gel. A 100bp Plus ladder (Fermentas) was included on the gel for determining band sizes. Following electrophoresis, the gels were visualised under UV light.

Table 2.1: Primer sets and sequence targets for RD analysis.

RD	Forward Primer	Chromosomal Position and Sequence Targets	Reverse Primer	Chromosomal Position and Sequence Targets	Reverse internal	Chromosomal Position and Sequence Targets
105	GTTCGTGCACA GTTGGGTG	**79424- 79406** CACCCAACTGTGCA CGAAC	ACCAGCTCCTC GACGCTATC	**83237-83256** GATAGCGTCGAGGA GCTGGT	GTTCAGTGCGC AGTTCGTTC	**79632-79651** GAACGAACTGCGCA CTGAAC
239	CCTGACCAGCA TCACTCCC	**4092026-4092008** GGGAGTGATGCTGG TCAGG	TCAAACCGTTCA CGACAAGC	**4092931-4092950** GCTTGTCGTGAACG GTTTGA	TCTACATCCCGA CGACCAGC	**4092660-4092641** GCTGGTCGTCGGGA TGTAGA
750	GTCAACTGCCG ATGGCTGAC	**1710600-1710581** GTCAGCCATCGGCA GTTGAC	GTGAACTAGGTC GAGCATCG	**1711722-1711741** CGATGCTCGACCTA GTTCAC	CGTCAGCGATGA TCACCTCG	**1711117-1711136** CGAGGTGATCATCG CTGACG
TbD1	GGGATTTCAG TGACTGGCCT G	**1761770-1761750** CAGGCCAGTCAC TGAAATCCC	TGTCCAAGGT TACGGTCACG C	**1761847-1761867** GCGTGACCGTAA CCTTGGACA	ACCGATAGAC GCTGAATCCC G	** **1745839-1745859** CGGGATTCAGCG

** Reference *Mycobacterium bovis* subsp. bovis AF2122/97 for TbD1.

2.3 Whole genome Next Generation Sequencing

2.3.1 Overview

A common application of the Illumina HiSeq 2000 platform is to sequence entire genomes of bacteria (or other organisms) using a paired-end library preparation approach. Whole genome sequencing on the Illumina HiSeq platform can give paired-end sequencing results. During the library preparation stage, the sample DNA is fragmented, and the fragments of a specific size (typically 200–500 bp, but can be larger) are ligated or “inserted” between 2 oligo adapters as depicted in Figure 2.1. The original sample DNA fragments are referred to as “inserts”. Therafter the DNA is amplified by PCR amplification and then sequenced from both ends of each fragment as indicated in Figure 2.2.

The raw files from the sequencing are in FASTQ format containing the reads with the corresponding quality value assigned to each base in a read. The read length depends on the library preparation and sequencing kits and method used. The paired-end reads obtained for the purpose of this study were 105 bp long. The entire insert from the adaptor is not sequenced as quality towards the end of a sequencing read diminishes (Cox *et al*., 2010).

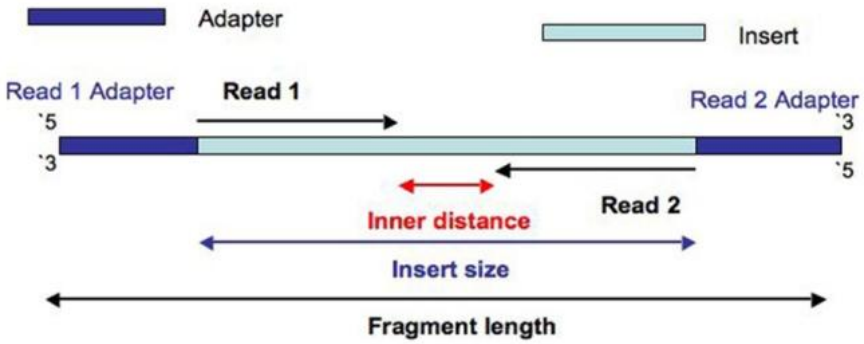


Figure 2.1: Diagram to show the construction of a fragment with an insert size is longer than the length of both reads *(Turner, 2014)*

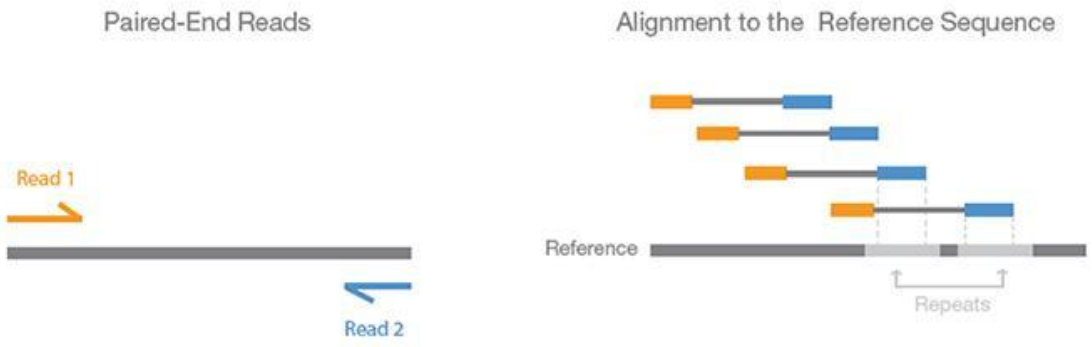


Figure 2.2: Paired-End Sequencing and Alignment: Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map reads over repetitive regions more precisely. This results in much better alignment of reads, especially across difficult to sequence, repetitive regions of genome (http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html).

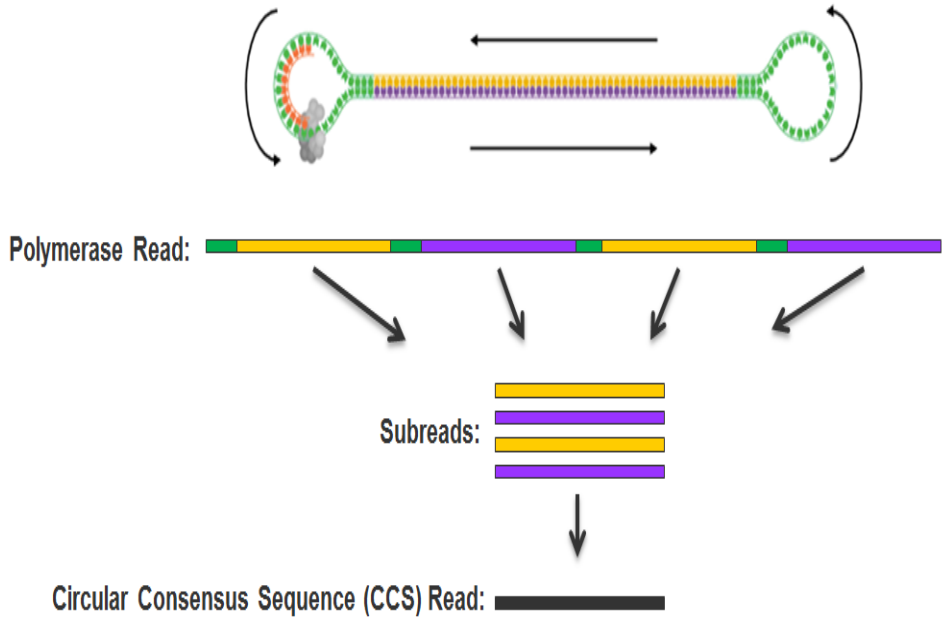


Figure 2.3: Read terminology for Pacbio reads illustrating how subreads and CCS reads are generated from a DNA template.

Subread: Each polymerase read is partitioned to form one or more subreads, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell™ template and no adapter sequences.

Read of insert: Represents the highest quality single sequence for an insert, regardless of the number of passes. For example, if your template received one-and-a-half subreads, that information will be combined into a Read of Insert. CCS is an example of a special case where at least two full subreads are collected for an insert. Reads of Insert give the most accurate estimate of the length of the insert sequence loaded onto a SMRT® Cell. For long templates, Reads of Insert may be the same as Polymerase Reads.

http://files.pacb.com/software/smrtanalysis/2.2.0/doc/smrtportal/help/!SSL!/Webhelp/Portal_PacBio_Glossary.htm

Long sequence reads can be obtained on the PacBio platform (PacBio RS) as a 'single molecule' sequence read as depicted in Figure 2.3. The raw reads from the PacBio RS sequencing are longer than the 105bp Illumina reads (median = 2,246 bp, maximum = 23,000 bp) but have a high error rate which must be corrected for (Keane *et al*., 2006; Tamura *et al*., 2011).

In this study, strains were sequenced on the Illumina HiSeq2000 platform giving paired 105bp sequence reads in FASTQ format. Additionally, 2 strains were also sequenced on the PacBio platform yielding sub-reads and circular consensus sequence (CCS) in FASTQ format as depicted in Figure 2.3.

2.4 NGS data analysis/Bioinformatics

2.4.1 Overview

The sequencing and library preparation for paired-end sequencing of the strains in this study was done in collaboration with Dr Arnab Pain and Dr Abdallah M. Abdallah from the King Abdullah University of Technology (KAUST), Saudi Arabia, and Dr Ruth McNerney and Dr Taane Clark from the London School of Hygiene and Tropical Medicine (LSHTM), UK. The sequencing platform used in this study was on the Illumina HiSeq2000 (Illumina, California, USA) platform. The sequences had a read length of 105bp and a base fragment size of 500 with an insert size ranging from between 350 and 550 bases.

The sequencing of 2 strains on the PacBio Platform was done in collaboration with Professor Alan Christoffels of the South African National Bioinformatics Institute (SANBI), University of the Western Cape, South Africa. Additional previously sequenced data for one of the strains analysed in this study was provided by Professor Dick van Soolingen and Dr Anita C. Schürch of Tuberculosis Reference Laboratory, National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Control, Bilthoven, The Netherlands.

Whole genome sequence mapping and alignment was done using only Illumina paired-end data whilst *de novo* assembly utilised both the Illumina and PacBio sequence data

2.4.2 FASTQ format

The FASTQ format is a text-based format which stores information for each read in terms of the nucleotide base sequence as well as the quality associated with each base assigned by the sequencer as illustrated in Figure 2.4. The first line, beginning with the '@'symbol, is called the 'header line' and corresponds to the name of the read and gives information such as the reverse or forward orientation of a read in the FASTQ file. The second line is the 'sequence line', representing the bases determined by a DNA

@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTT
CCAAA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#"""""""""""""""7F@71,'";C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?

Figure 2.4: FASTQ Format Description: The first line, beginning with the '@'symbol, corresponds to the title of the FASTQFASTQ file and is called the 'header line'. The second line is the 'sequence line', representing the bases determined by a sequencing machine. An optional tile line beginning with the '+'sign comes below the sequence line and this is then followed by a 'quality line'that corresponds to the 'sequence line'. It follows from this that 'sequence line' and 'quality line'are the same length (Cock *et al*., 2010).

sequencer. An optional line beginning with the '+'sign comes below the sequence line and

this is then followed by a 'quality line' that corresponds to the 'sequence line', assigning a quality value to each base in the sequence line. It follows from this that 'sequence line' and 'quality line'are the same length.

2.4.3 Quality Score

The quality score is given as a Phred score which is a probability of a base being called correctly by a sequencer: $Q_{Phred} = -10 \times \log_{10}(P_e)$.

The quality scores (quality line characters shown in Figure 4) are encoded as ASCII (American standard code for informational change) printable characters which correspond to the ranges illustrated in Table 2.2. The Phred quality scores corresponding base call accuracies are illustrated in Table 2.3. The system correlates a character with a number. For example, the character "=" represents a Phred quality score of "28", that correlates to an error probability of 0.00158.

Table 2.2: Sanger FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores.

Description , **OBF name	**ASCII characters**		**Quality Score**	
	Range	**Offset**	**Type**	**Offset**
Sanger standard fastq-sanger	33-126	33	Phred	0 to 93

(Cock *et al*., 2010)

** Open Bioinformatic Foundation (OBF)Projects FASTQ key variants, and conventions adopted by the Open Bioinformatics Foundation (OBF, http://www.open-bio.org).

Table 2.3: Phred quality scores are logarithmically linked to error probabilities.

Phred Quality Score	**Probability of incorrect base call**	**Base call accuracy**
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

(http://en.wikipedia.org/wiki/Phred_quality_score)

2.4.4 Pre-Processing Sequence Reads

The quality of the FASTQ sequence reads was assessed using the FASTQC program (Andrews, 2010).

In this study, a cut-off mean Phred score value of ≥20 (corresponding to a base call accuracy of ≥99%) was taken for further processing of reads. Paired end sequence reads

with a mean quality score of less than 20 were trimmed using Fastx Toolkit so as to have a mean quality of 20 (Hannon Lab, 2009). FASTA/Q Trimmer version 0.0.6 from Fastx Toolkit was used from command line to trim as shown in Figure 2.5. The 3' ends of the reads were trimmed for low quality bases with fastx-trimmer and the FASTQC results were used to infer how many bases were trimmed from each read.

```
Commandlineusage: fastx_trimmer [-f N] [-l N] [-v] [-i INFILE] [-o OUTFILE]

[-f N]          = First base to keep. Default is 1 (=first base).
[-l N]          = Last base to keep. Default is entire read.
[-i INFILE]     = FASTA/Q input file. Default is STDIN.
[-o OUTFILE]    = FASTA/Q output file. Default is STDOUT
```

Figure 2.5: the linux command line usage for the trimming of fastq files using fastx toolkit. The command is run without the inclusion of the square brackets '[ ]' which indicate what variables to input with the options '-f', '-l', '-i' and '-o' (hannon lab).

PacBio long sub-reads, with inherent errors, were error corrected using the shorter length, high fidelity sequences or circular consensus sequence (CCS) provided with the sub-reads. The sub-reads and CCS reads are derived from the same template as mentioned earlier and illustrated in Figure 2.3. CELERA read correction software for PacBio reads can be used for error correction (Denisov *et al*., 2008; Miller *et al*., 2008). The CELERA Assembler however requires the converting of "FASTQ" files from the sequencing machines into "frg" format. The high fidelity CCS FASTQ files are used to error correct the longer error riddled single long reads as shown in Figure 2.6. If CCS FASTQ files are not provided, Illumina paired-end reads can be used for error correction as long as they are derived from the same DNA template used to generate the PacBio long reads. In this study high fidelity PacBio CCS reads were used to error correct the long, error riddled sub-reads using CELERA read correction software.

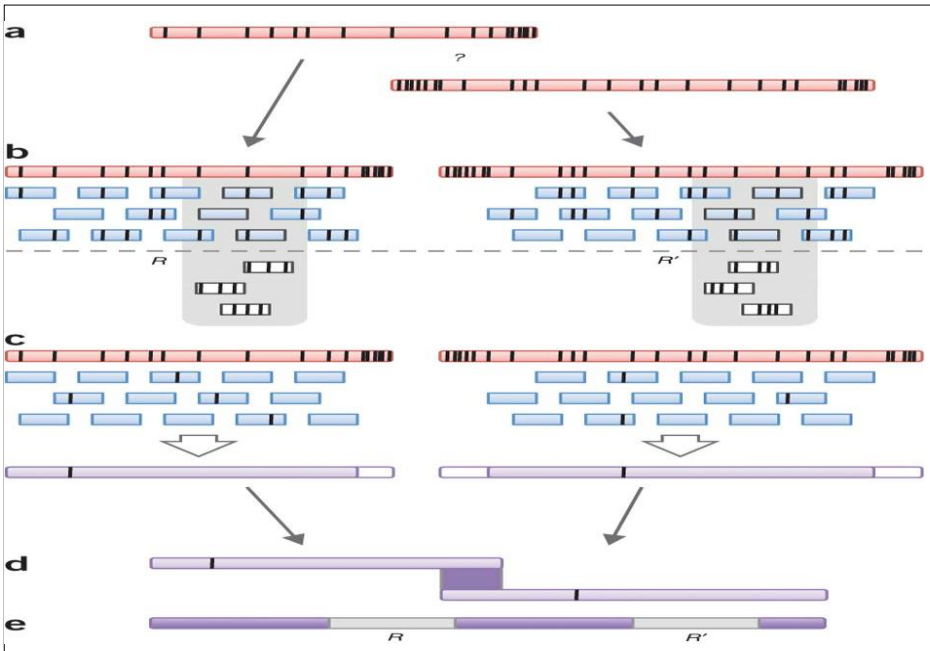


Figure 2.6: Errors, indicated by black vertical bars, in single-pass PacBio RS reads (pink rectangles) make it difficult to determine whether reads overlap. (b) Aligning high-fidelity short reads to error-prone long reads. Accurate alignments can be computed because the error between a short, high-accuracy sequence (~99% identical to the truth) and a PacBio RS sequence is half the error between two PacBio RS sequences. In this example, black bars in the short-reads indicate 'mapping errors' that are a combination of the sequencing error in both the long and short reads. In addition, a two-copy inexact repeat is present (outlined in grey) leading to pile-ups of reads at each copy. To avoid mapping reads to the wrong repeat copy, the algorithm selects a cut-off, C, and only the top C hits for each short read are used. The spurious mappings (in white) are discarded. (c) The remaining alignments are used to generate a new consensus sequence (purple), trimming and splitting long reads whenever there is a gap in the short-read tiling. Sequencing errors, indicated in black, may propagate to the PacBio corrected read (PBcR) in rare cases where sequencing error co-occurs. (d) After correction, overlaps between long PBcR sequences can be easily detected. (e) The resulting assembly is able to span repeats that are unresolvable using only the short reads. (Koren *et al*., 2012)

2.4.5 Mapping of Sequence Reads to a Reference Genome

Mapping software is used to map/align the reads contained in FASTQ files to a reference genome (*M. tuberculosis* H37Rv). H37Rv was used as a reference so as to be able to compare our results to those of others as has been common practice (Coll *et al*., 2014; Comas *et al*., 2013; Luo *et al*., 2015; Merker *et al*., 2015; Schürch *et al*., 2011a, 2011b). The use of a Beijing reference would have however been useful as regions which are absent in H37Rv would have been better analysed. Mapping of reads, via the command line can be done using various alignment algorithms such as BWA (http://bio-bwa.sourceforge.net/) (Li and Durbin, 2009), BFAST (Homer *et al*., 2009), NOVOALIGN (http://www.novocraft.com/support/download/)(Novocraft) and SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) (Hannes Ponstigl, 2014). In this study, we made use of BWA, SMALT, and NOVOALIGN to map FASTQ reads to the reference H37Rv as is illustrated in Figure 2.7. The information of the mapping is stored in the Sequence Alignment/Map (SAM) format (Li *et al*., 2009). Mapping algorithms can broadly be categorized as either hash table based or Burrows Wheeler transformation (BWT)-based (Hatem *et al*., 2013). BWA uses an index built with the BWT whilst Novoalign hash table is built by dividing the reads into overlapping oligomers and uses the Needleman-Wunsch algorithm for alignment (Hatem *et al*., 2013; Ruffalo *et al*., 2011). SMALT also uses a hash table for mapping to a reference genome using dynamic programming (http://www.sanger.ac.uk/science/tools/smalt-0). For the hash-table, common sub-strings of characters between the reference genome and the reads is sought by a process called seed detection of which different strategies among aligners result in aligner differences (Shang *et al*., 2014)

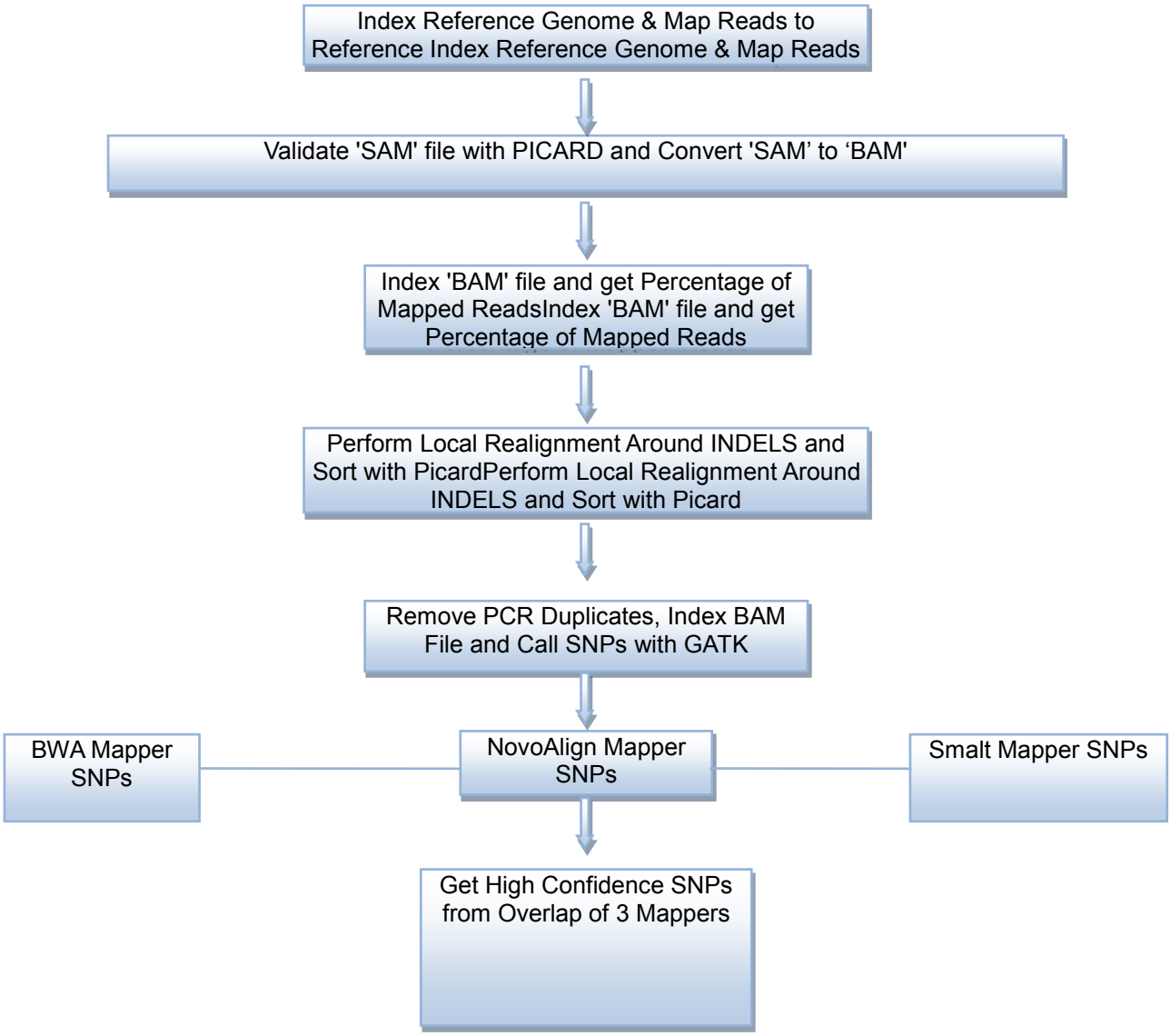


Figure 2.7: Flow diagram of whole genome mapping using 3 mapping software algorithms.

2.4.6 SAM Format

The SAM file consists of a header section, containing various information fields, and the read sequence with the corresponding mapping quality for each base (in contrast to the FASTQ file, where that quality of each base refers to the probability of the sequencer to have called/placed that base correctly). Each header line record begins with an "@" and has various tags under it. The alignment information field gives information about how the reads from the sequencer mapped to the reference. This includes flags for the quality of mapping, number and pairs of reads mapping as well as the coordinates and orientation of the reads with respect to the genome reference mapped to. Information found in the alignment section is depicted in Figure 2.8 and Table 2.4. The SAM file, which is in text format, can be subsequently converted to its binary format, the BAM file (Aabye *et al*., 2011; Li *et al*., 2009).

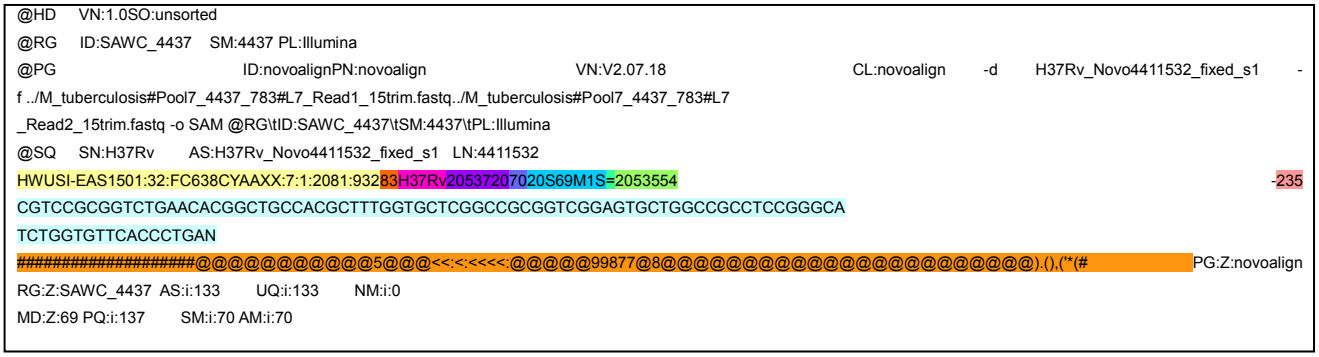


Figure 2.8: Sequence alignment in SAM format. Highlighted portions are described in Table 3 according to highlighting colour.

Table 2.4: Description of tags found in SAM file sequence alignment.

Tag	Description	
@HD The header line. The first line if present		
VN	Format version. Accepted format: /^[0-9]+\.[0-9]+$/	
SO	Sorting order of alignments	
@SQ Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order		
SN	Reference sequence name. Each @SQ line must have a unique SN tag	
LN	Reference sequence length	
AS	Genome assembly identifier	
@RG Read group		
ID	Read group identifier. Each @RG line must have a unique ID. The value of ID is used in the RG tags of alignment records. Must be unique among all read groups in header section	
SM	Sample. Use pool name where a pool is being sequenced	
@PG Program		
ID	Program record identifier. Each @PG line must have a unique ID.	
PN	Program name	
CL	Command line	
Other Tags		
Tag	**Type**	**Description**
AS	i	Alignment score generated by aligner
MD	Z	String for mismatching positions. Regex : [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*7
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping
PG	Z	Program. Value matches the header PG-ID tag if @PG is present
RG	Z	Read group. Value matches the header RG-ID tag if @RG is present in the header
AM	i	The smallest template-independent mapping quality of segments in the rest
SM	i	Template-independent mapping quality
UQ	i	Phred likelihood of the segment, conditional on the mapping being correct

Table 2.5: sequence alignment/map (sam) format is tab-delimited. Apart from the header lines, which are started with the '@' symbol, each alignment line consists of:

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string, provides information as to how the read aligned to the reference genome, e.g. M = match, , I = insertion compared to reference, D = deletion compared to the reference
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)

Mapping of sequence reads to a reference genome typically involves the indexing of the reference genome. This comprises first dividing the reference genome sequence into units or word lengths of a particular size (k-mers). The indexing step size, s, is then set where if s=1, every k-mer is indexed; s=2, every second k-mer is indexed and so on. Following the indexing of the reference genome, the FASTQ files containing the reads of the sequenced genomes are then mapped to the reference genome according to respective user guides of the software. The alignment is produced in SAM format.

In this study, BWA, NOVOALIGN and SMALT were used to map FASTQ files from the sequenced genomes to the *M. tuberculosis* H37Rv (-emb|AL123456.3|) (http://www.ncbi.nlm.nih.gov/nucleotide/444893469?report=genbank&log$=nuclalign&blast_rank=4&RID=K91P4D1W016#_blank) reference genome.

2.4.7 Processing the SAM and BAM Files

The SAM and BAM files generated by sequence mapping software can be improved upon and made compatible with downstream processing programs by utilising SAM and BAM processing tools. These include SAMtools (http://samtools.sourceforge.net/), PICARD (https://github.com/broadinstitute/picard), GATK (https://www.broadinstitute.org/gatk/) and BEDTOOLS (http://bedtools.readthedocs.org/en/latest/) (Quinlan, 2014; Quinlan and Hall, 2010).

2.4.7.1 SAMTOOLS

SAMtools can be used to determine alignment statistics from the SAM files, convert SAM files to BAM files, index BAM files for quick retrieval of alignment information, and other manipulations of the SAM/BAM files, as well as SNP and in/del detection.

2.4.7.2 Genome Analysis Toolkit (GATK) and PICARD

A number of post-processing of the SAM and BAM files generated was done using GATK and PICARD. These comprised the following:

- Validate that the SAM file has been correctly constructed by the genome mapping software.
- Add read groups to the SAM/BAM files that may be required for downstream analysis.
- Sort the BAM file to ensure quick retrieval of mapping information.

- Mark and remove PCR duplicates from an alignment – if multiple read pairs have identical external coordinates upon being mapped to a reference, it is best to keep the read pair with the highest mapping quality (Li *et al*., 2009).
- Local realignment around indels - serves to correct for misalignment of reads to the reference genome as a result of the presence of an insertion or deletion with respect to the reference. The process involves the identification of such indels and then realigning reads at these sites.
- Call SNPs from BAM files.

2.4.8 SNP analysis

GATK was used to identify SNPs from the sequence alignment (BAM) files and an in-house Python script was used to do the following:

- Annotate SNPs as being synonymous, non-synonymous or non-coding.
- Report SNPs by "functional category".
- Identify and calculate the number of high confidence SNPs from the overlap of 3 mapping algorithms (BWA, SMALT and NOVOALIGN).

In order to test the validity of our high confidence SNP calling approach, we compared SNP positions previously verified in two samples (SAWC 507 and SAWC 5527) by Schürch *et al*., 2011b to that called by the high confidence SNP calling approach used here on the same two re-sequenced strains verified at 273 positions. SNP identities of our high confidence SNP sets at the aforementioned positions were compared to the previously verified positions so as to test the accuracy of our high confidence SNPs.

An in-house Python script was used to concatenate SNPs called by GATK for each genome and these were used as alignment files as was done by Schürch *et al*., 2011b.

2.4.9 Phylogenomic Analysis

2.4.9.1 Identification of informative SNPs

The construction of phylogenetic trees to infer the evolutionary history of the Beijing lineage family of *M. tuberculosis* was done using concatenated SNPs generated for each strain studied. The concatenated SNPs used were first analysed to determine whether these SNPs were able to distinguish one strain from the other and also

identify which SNPs were informative in the discrimination. In this study SINGLE NUCLEOTIDE POLYMORPHISM TYPING tool (SNPT) (http://www.shigatox.net/stec/cgi-bin/snpt) (Filliol *et al*., 2006) was used to determine the discriminatory index of concatenated SNP sequence types where an index of 1 indicates sequences analysed can be discriminated 100%. Additionally, the set of informative SNPs that were informative in the discrimination was also calculated.

2.4.9.2 Construction and congruency of phylogenetic trees

Phylogenetic trees can be used to depict the evolutionary scenario of a set of analysed species as well as show the evolutionary position of the common ancestors in such analysis (Schürch *et al*., 2011b). The construction of phylogenetic trees requires a substitution model on which to base the construction of a phylogenetic tree.

In this study MODELGENERATOR (http://mcinerneylab.com/software/modelgenerator/#) was used to calculate the most appropriate model to use in phylogenetic tree construction (Keane *et al*., 2006; Tamura *et al*., 2011).

Following the identification of an appropriate model to use for construction of phylogenetic trees, SEAVIEW genome analysis software was used to construct maximum likelihood, neighbour-joining and phyml maximum likelihood phylogenetic tree (http://doua.prabi.fr/software/seaview) (Gouy *et al*., 2010). The substitution model to use in the phylogenetic tree construction was based on the best-fit model as determined by MODELGENERATOR and 1,000 bootstrap pseudo replicates were applied in the tree building. Phylogenetic trees constructed from genome-wide SNPs were subsequently comparatively analysed to phylogenetic trees based on:

- insertion sequence points plus selected 3 SNPs in 3R genes (Hanekom *et al*., 2007a).
- 41 3R System SNPs (Mestre *et al*., 2011).
- Genome-wide sSNP.
- Genome-wide nsSNPs.

To calculate the correlation of phylogenetic trees constructed, unrooted phylogenetic trees which were saved in SEAVIEW were opened using FIGTREE (http://tree.bio.ed.ac.uk/software/figtree/) and saved in nexus ('NEX') format to

enable input into MESQUITE software (mesquiteproject.wikispaces.com). MESQUITE was subsequently used to measure congruency of trees generated using the 'Patristic Distance Correlation' as a measure. Values of correlation range from zero to one with a value of 1 indicative of 100% congruency (Maddison W.P. Maddison D.R., 2015).

2.4.9.3 SNP Analysis based on the topology of phylogenetic trees

The phylogenetic trees constructed provided an illustration of the evolutionary analysis of the *M. tuberculosis* Beijing strains analysed in this study. This information aided the identification of SNPs that were unique to a sub-lineage or common to a set of lineages sharing a common node on phylogenetic trees constructed. The identification of unique and common SNPs was achieved through an in house Python script written by Dr Ruben Gerhard van der Merwe. Further analysis for high confidence unique and common SNPs among lineages was done in MICROSOFT EXCEL. In addition to the analysis of identification of high confidence SNPs in MICROSOFT EXCEL, the relative position of non-coding SNPs was determined and compared to previously identified transcriptional start sites (TSS) and the start position of the nearest gene in correct orientation to the non-coding SNP (Cortes *et al*., 2013; Forse *et al*., 2011; Rose *et al*., 2013). The nsSNPs, were further analysed for gene ontology biological processes uniqueness using PANTHER (protein annotation through evolutionary relationship) classification system (http://www.pantherdb.org/) (Mi *et al*., 2013). This involved pasting the gene harbouring the SNPs according to TubercuList nomenclature (http://tuberculist.epfl.ch/index.html) into the PANTHER web interface (http://pantherdb.org/).

2.4.10 Analysis of areas with zero- and more than double sequence read mapping for identifying putative deletions and duplications

Large genomic deletions (or regions of difference) have been identified in the genomes of *M. tuberculosis* strains and have been used as genetic markers and also implicated in virulence in some studies. These regions of difference have been identified with respect to the reference strain H37Rv. Subsequently in NGS whole genome mapping using H37Rv, regions of the reference which are absent in the query strains being analysed result in there being zero coverage in the mapping of the query strains reads to a reference genome.

2.4.10.1 BEDTOOLS

BEDTOOLS is a software suite that allows for the interrogation of genomic DNA sequence for a variety of features. Included among its capabilities is the ability to calculate genome coverage of aligned sequence reads to a reference genome (Quinlan, 2014; Quinlan and Hall, 2010).

Bedtools was used to calculate the areas of zero depth of coverage (no reads aligned to the reference genome) so as to identify deletions in the sequenced genomes with respect to *M. tuberculosis* H37Rv. Additionally, areas that had more than 1.8x the mean number of mapped reads to the *M. tuberculosis* H37Rv reference genome were identified as possible sequence duplications. Regions of zero genome coverage were compared to known deletions previously associated with *M. tuberculosis* Beijing strains. Uncharacterised areas of zero coverage were reported as potential large deletions.

2.4.10.2Primer design for novel deletions

For the design of primers for detection of large deletions, 500bp of sequence up and downstream of the deletion was selected in addition to the deleted sequence.

For detection of genomic sequence not deleted, primers were designed such that the 'Forward Primer' was in the 500bp region upstream of the deleted sequence, the 'Reverse Internal Primer' was in the potentially deleted sequence Figure 2.9. Presence of a PCR product whose size is determined by combination of the Forward

and Reverse Internal primers implies that the region is not deleted in a strain being tested.

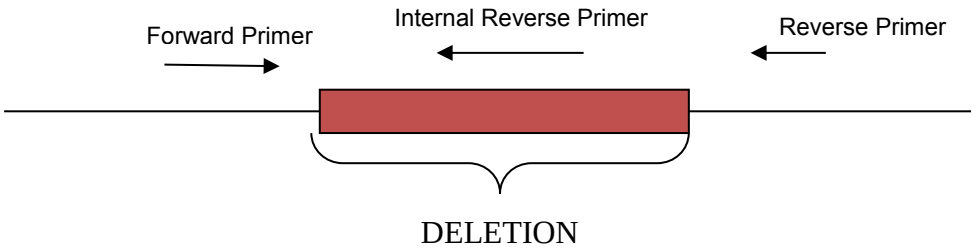


Figure 2.9: Design and placement of primers to detect both deletion and non-deletion events. A product size based on forward-reverse primers implies a deletion even whilst a product size based on forward-reverse internal primers implies a non-deletion event.

For detection of deletions, primers were designed such that the 'Forward Primer' was in the 500bp region upstream of the deleted sequence and the 'Reverse Primer' was in the 500bp region downstream of the deleted sequence as depicted in Figure 2.8. Presence of a PCR product whose size is based on the Forward and Reverse primers implies that the region is not deleted in a strain being tested.

2.4.11 Genome Assembly of Sequence Reads

Sequence reads can also be put together without the use of a reference genome through a process called *de novo* assembly. The overlap between reads is used to build a continuous stretch of sequence called a contig as depicted in Figure 2.9. Various software tools have been developed to carry out *de novo* assembly and include MIRA and CELERA algorithms (Baker, 2012; Chevreux *et al*., 2004; Denisov *et al*., 2008; Kajitani *et al*., 2014).

2.4.12 Pre-processing and genome assembly

A number of genome assembly software tools are able to do *de novo* assembly utilizing sequences derived from two different platforms (e.g. Illumina P.E. and PacBio reads) or sequencing libraries, resulting in a hybrid assembly. Hybrid assemblies have been shown to be of higher quality than assemblies generated from reads from a single platform (Koren *et al*., 2012; Liao *et al*., 2015). Two software packages capable of hybrid assembly of PacBio and Illumina paired-end sequences are CELERA and MIRA (Chevreux *et al*., 2004; Denisov *et al*., 2008; Koren *et al*., 2012). CELERA uses a dynamic window approach where reads are aligned to

confirm bases followed by the merging of aligned reads with common bases using CABOG (Celera Assembler with Best Overlap Graph) whereas for MIRA every read is compared to every other read for overlap followed by local alignment using Smith-Waterman algorithm (Chevreux *et al*., 1999, 2004; Denisov *et al*., 2008; Miller *et al*., 2008, 2010).

Pre-processing of reads for genome assembly using CELERA and MIRA are slightly different. The CELERA assembler requires that all input files be converted from "FASTQ" files into "FRG" format. MIRA on the other hand requires Illumina FASTQ files not be pre-processed. MIRA requires that PacBio input reads to be error corrected as is the case for CELERA. Another common feature of CELERA and MIRA is that they both require that a parameter file be set up which also states where input and output files are to be placed and how the assembly is to be run. This is called the "spec file" in CELERA and "config file"in MIRA. The assembled genome comprises contigs which are a consensus stretch of overlapping reads. A number of contigs can result from a single assembly as a result of failure to determine precisely the overlap of some sequences. This is usually due to 'repeat' sequences. An illustration of genome assembly is given in Figure 2.10 which shows how reads that overlap in sequence are assembled into contigs and subsequently into scaffolds.

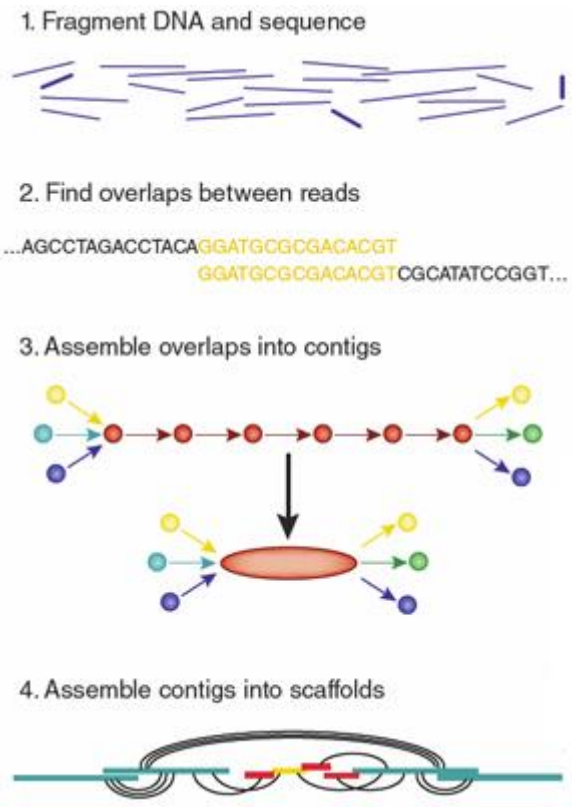


Figure 2.10: Genome assembly stitches together a genome from short sequenced pieces of DNA (Michael Schatz, Cold Spring Harbor) (Baker, 2012).

2.4.13 CELERA assembly

CELERA is a whole genome sequencing assembler that is capable of hybrid assembly involving DNA sequences from different platforms. In this study we used CELERA programs for hybrid genome assembly as illustrated in Figure 2.11 using Illumina paired-end and PacBio sequence data.

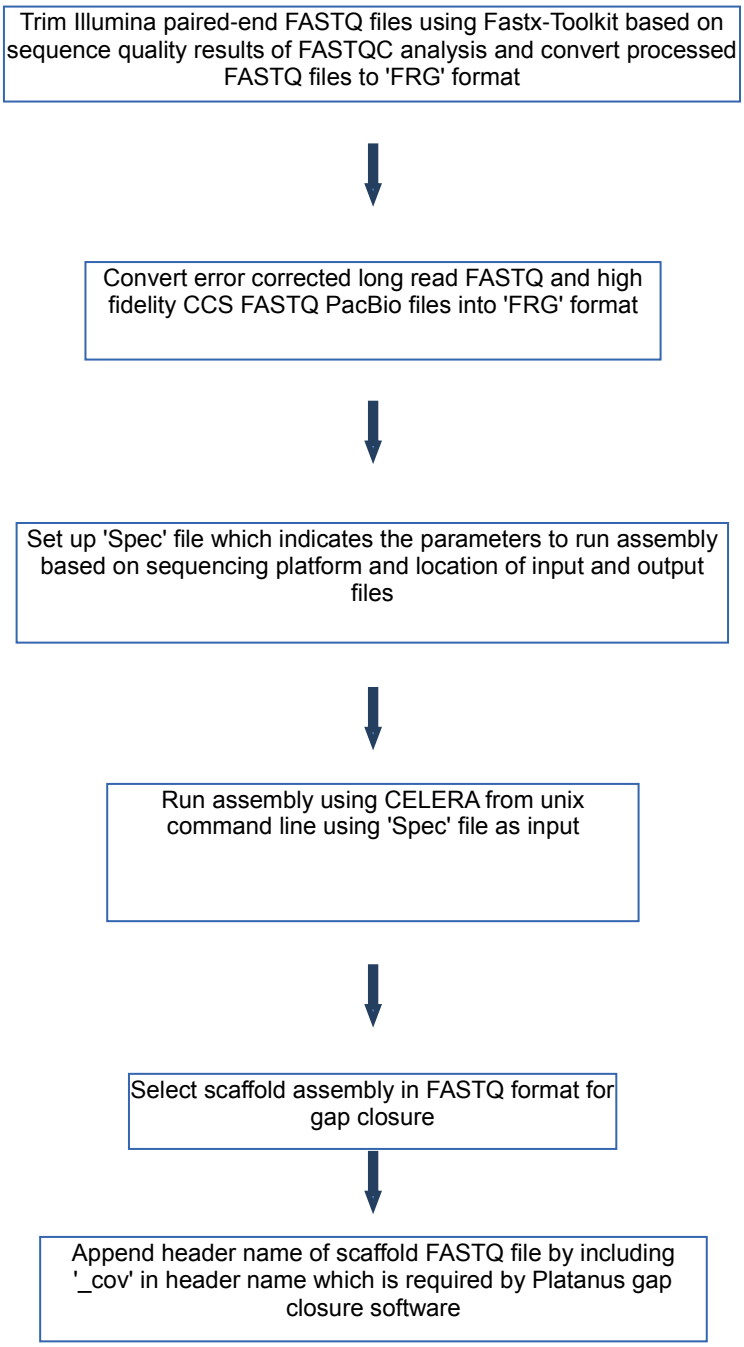


Figure 2.11: Flow diagram of whole genome assembly using CELERA software.

2.4.14 MIRA Assembly

MIRA is a multi-pass whole genome mapper and assembler that is capable of utilizing sequence data from different platforms in a hybrid assembly. Supported platforms include Illumina and PacBio sequence data. The use of MIRA in this study is outlined in Figure 2.12 and involves the writing of a manifest configuration file which contains information on the parameters of running the assembly and which sequence data to load for the assembly.

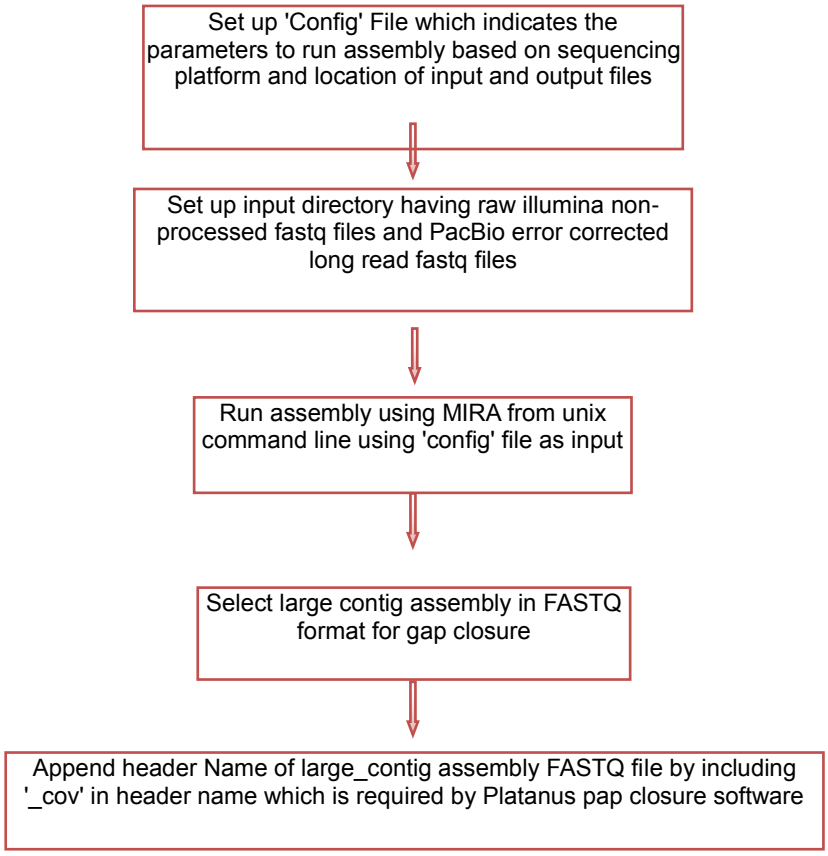


Figure 2.12: Flow diagram of whole genome assembly using MIRA software.

2.4.15 GAP closure of the assembled genome

The non-overlap of contigs can be resolved by making use of paired-end sequencing information. When one pair of a read is found on a particular contig and its partner is found on another, the 2 contigs can be joined together. This is particularly useful if the insert size of the read pair spans a repeat element or has part of its sequence overlapping a repeat element as illustrated in Figure 2.13a and Figure 2.13b. Information of the insert size is what enables the joining of contigs in the aforementioned case.

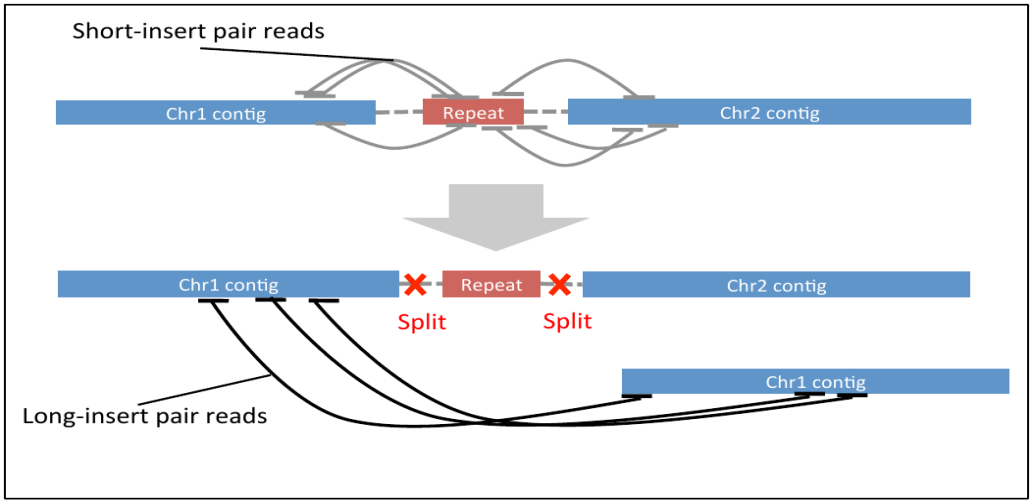


Figure 2.13a: Paired reads resolving a gap between to contigs with repeat sequences within them as a result of at least one pair of the reads not overlapping the repeat sequence.

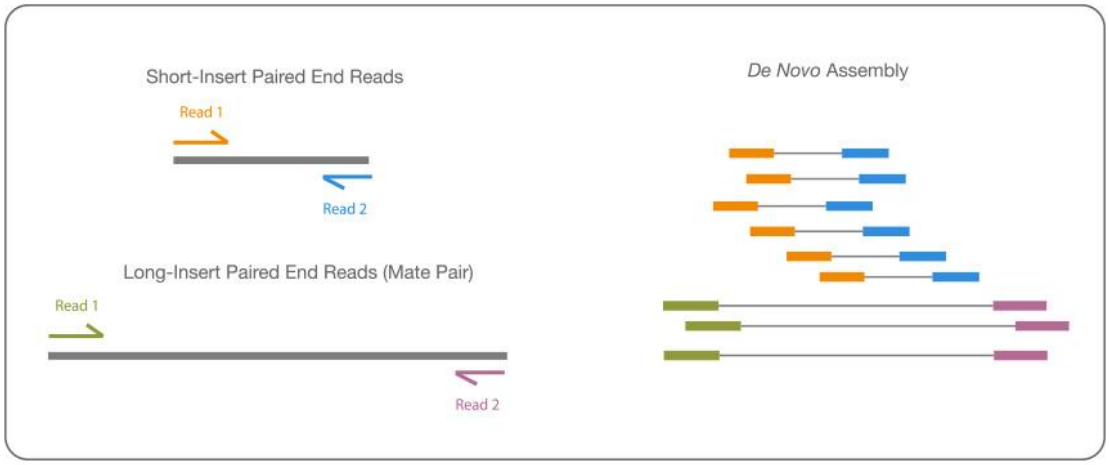


Figure 2.13b: Combining short-insert, paired-end and long-insert, mate pair sequences is the ideal way to maximize coverage as long insert mate pairs can bring 2 contigs spanning a repeat whilst paired end reads with short insert size can fill in the genome.

In this study PLATANUS assembler (http://platanus.bio.titech.ac.jp/) was used to order and close gaps in genome assemblies using Illumina paired-end reads (Kajitani *et al*., 2014). This was followed by the aligning and ordering of contigs and super-contigs produced by *de novo* assembly to a reference genome using ABACAS. This aids in visualization of assembled genomes as well as subsequent comparative analysis based on synteny (Assefa *et al*., 2009; Swain *et al*., 2012). The visualization of how well a genome was assembled when compared to a reference genomes and subsequent comparative analysis was undertaken using PROGRESSIVE MAUVE (http://darlinglab.org/mauve/mauve.html). The less the number of collinear blocks, the better a genome assembly when compared to a reference or between two or more assembled genomes (Darling *et al*., 2010).

2.4.16 Identification and location of IS6110 sequence in assembled genomes

The identification and location of the IS*6110* element in the assembled genomes was done as manual search process of the IS*6110* element in the genome assemblies and blasting adjacent sequence to the IS*6110* element match in the genome using NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome) (Altschul *et al*., 1997). The process for identification and location of the IS*6110* in assembled genomes in this study is illustrated in Figure 2.14.

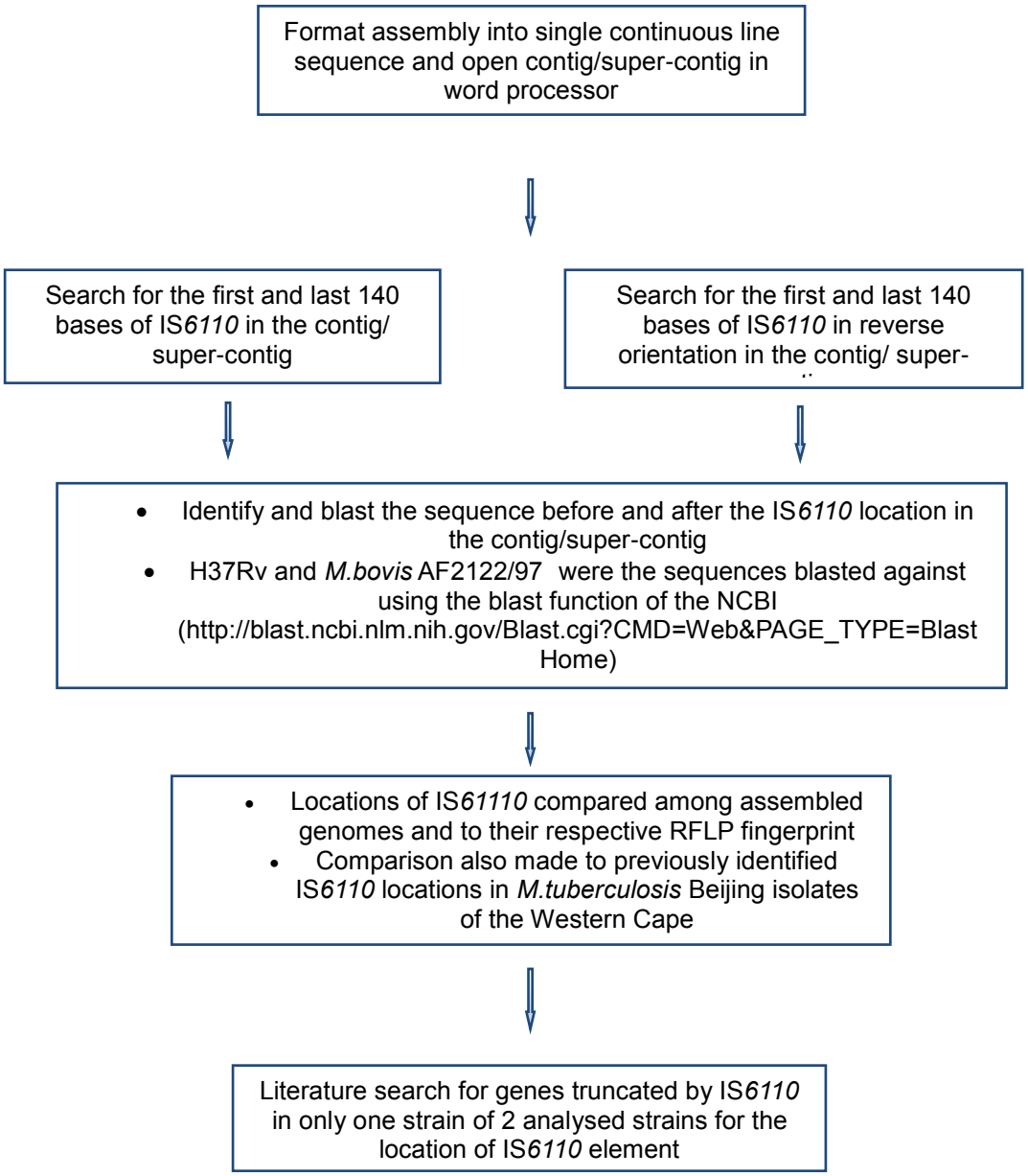


Figure 2.14: Flow diagram for the identification of the IS*6110* element in assembled genomes.

2.4.17 *Identification of tandem repeats*

An analysis of repeats in the assembled hyper-hypo virulent genomes was undertaken using JEMBOS (http://emboss.sourceforge.net/Jemboss ETANDEM repeat finder. Additionaly, a manual search of MIRU/VNTR repeats proposed by Supply *et al*., (2006) was done in assembled genomes

3 RESULTS

3.1 Molecular methods

Spoligotyping of the selected strains in this study were that of the classical Beijing genotype for all strains previously described as Beijing. The strains also had the RD105 and TbD1 deletions but had intact RD750 and RD239 deletions.

3.2 Mapping of sequence reads to a reference genome

Whole genome mapping was performed using 3 different genome mapping algorithms namely NOVOALIGN, BWA and SMALT. Quality processed paired-end Illumina fastq files were used as input for the genome mapping and were aligned to the reference genome, *M. tuberculosis* H37Rv. The output mapping format was a SAM file (previously described in the materials and methods section) for all mapping algorithms. All SAM files generated were validated with PICARD software that they were correctly generated and subsequently converted to the BAM format. Following the generation of the BAM files, duplicate reads were removed from the BAM file so as to only retain reads with the highest quality. Additionally, reads were realigned around small insertions, duplications and deletions to rectify misalignments of reads to the reference which can lead to spurious SNP callings.

An example of the SAWC 507 statistics of the alignment for the final BAM files generated by the 3 different mapping algorithms are shown in Table 4.1 as determined by QUALIMAP (García-Alcalde *et al*., 2012). Results of alignment statistics of other strains in this study are given in Supplemental data.

Table 3.1: Mapping statistics of SAWC 507.

Strains	Mapper	Number of reads	Mapped reads (%)	Unmapped reads (%)	Paired reads (%)	Mean Mapping Quality
SAWC 507	NovoAlign	36,867,796	35,423,666 (96.08%)	1,444,130 (3.92%)	35,423,666 (96.08%)	69.68
SAWC 507	BWA	23,738,244	23,385,078 (98.51%)	353,166 (1.49%)	23,385,078 (98.51%)	55.74
SAWC 507	SMALT	23,984,104	23,786,669 (99.18%)	197,435 (0.82%)	23,786,669 (99.18%)	43.71

The mapping statistics showed that there was some variation in the manner in which

the mapping algorithms aligned the sequence reads to the reference genome. Repetitive regions are a contributing factor and this can have an effect on downstream analysis such as SNP calling and identification of large deletions. Mapping of sequence reads in some repeat areas such as PE-PGRS genes was observed to drop. However there were a significant number of repeat areas where the mapping was similar to non-repeat areas of the genome.

3.3 SNP calling

Genomic differences between whole genome sequenced organisms can be compared by identifying SNPs relative to a reference genome. In this study our sequenced strains were aligned to the reference genome *M. tuberculosis* H37Rv using 3 different mapping algorithms which had slight differences in their mapping statistics as indicated in Table 4.1. To circumvent this, high confidence SNPs, which we described as being common to all 3 mapping algorithms, were sought. The tool used to call SNPs from the processed BAM alignment files was GATK. Results of the SNP calling showed that there was variation in the number of SNPs which were called for a sequenced genome aligned by 3 different mappers. An example of this is shown in Figure 3.1 for strain SAWC 507 and this occurs in all the strains analysed. The overlap is considered high confidence, and used in subsequent analyses. The number of high confidence SNPs for all the strains in this study are given in Supplemental data.

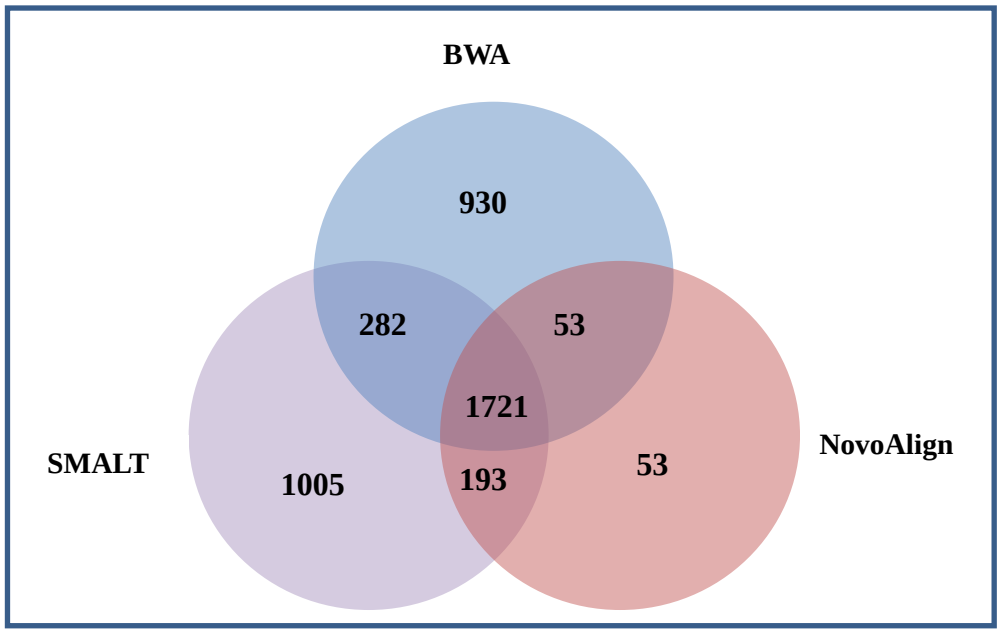


Figure 3.1: SNPs which were called for a sequenced genome aligned by 3 different mappers and showing their overlap.

To validate the methodology of identification of high confidence SNPs, we compared 273 SNP positions in SAWC 5527 and SAWC 507 strains that had previously been verified in a study by Schürch *et al*. (2011). There was agreement in SNP calling at 272 positions out of 273 positions.

3.4 Comparative whole genome SNP analysis based on 7 sub-lineages

Coding and non-coding genome-wide SNPs were compared among the representative strains of the 7 sub-lineages of Beijing in this study. A total of 929 SNPs were common to all the representatives of sub-lineages with respect to the reference, *M. tuberculosis* H37Rv. A more focused analysis into the typical and atypical Beijing groupings revealed that the atypical Beijing strains had 943 common SNPs and 10 SNPs were common amongst the atypical strains but not found in the typical Beijing strains. On the other hand, the typical Beijing strains had 1277 common SNPs of which 348 SNPs were absent in the atypical Beijing grouping (i.e. unique to the typical Beijing strains).

3.5 Comparison to previously described evolutionary studies

3.5.1 Approach

The evolutionary analysis of members of the 7 Beijing sub-lineages was done using a whole genome sequencing approach which is the gold standard for phylogenetic analysis. Concatenated high-confidence SNPs from the Beijing trains, reference genome *M. tuberculosis* H37Rv and a sequenced *Mycobacterium bovis* isolate were subsequently aligned to generate phylogeny input file.

Phylogenetic trees were constructed using SEAVIEW (Gouy *et al*., 2010) to predict the evolutionary relationship between the Beijing strains. These trees were generated using either the maximum likelihood, parsimony or neighbour-joining algorithms.

Trees based on Maximum Likelihood, Parsimony and Neighbour-Joining were compared to establish congruence in order to determine confidence in the tree topology. In addition, the trees were compared to the evolutionary scenario described by (Hanekom *et al*., 2007b) which was based on *mutT* genes, IS*6110* insertion sites and previously described sSNPs and nsSNPs. Finally, our trees were compared to the trees based on verified SNPs and RDs according to (Schürch *et al*., 2011b).

3.6 Identification of informative SNPs

The construction of phylogeny trees using only sSNPs enables the description of sub-lineage evolution on the basis of being less likely to be under selective pressure (Coll *et al*., 2014) (Coll *et al.,* 2014).The genome-wide sSNPs discriminatory index was 1 and was associated with 288 informative SNPs when assessed using SNPT (www.shigatox.net/stec/cgi-bin/snpt).

3.7 Phylogenetic trees based on full set of genome-wide SNPs

The initial approach of the phylogenetic analysis in this study involved using all the high confidence SNPs amongst our strains. Figures 3.2 and 3.3 illustrate the phylogenetic analysis of the Beijing strains in this study using high confidence SNPs according to the maximum parsimony and maximum likelihood methods. The majority rule was used to support nodes using bootstrap values from 1000 replicates.

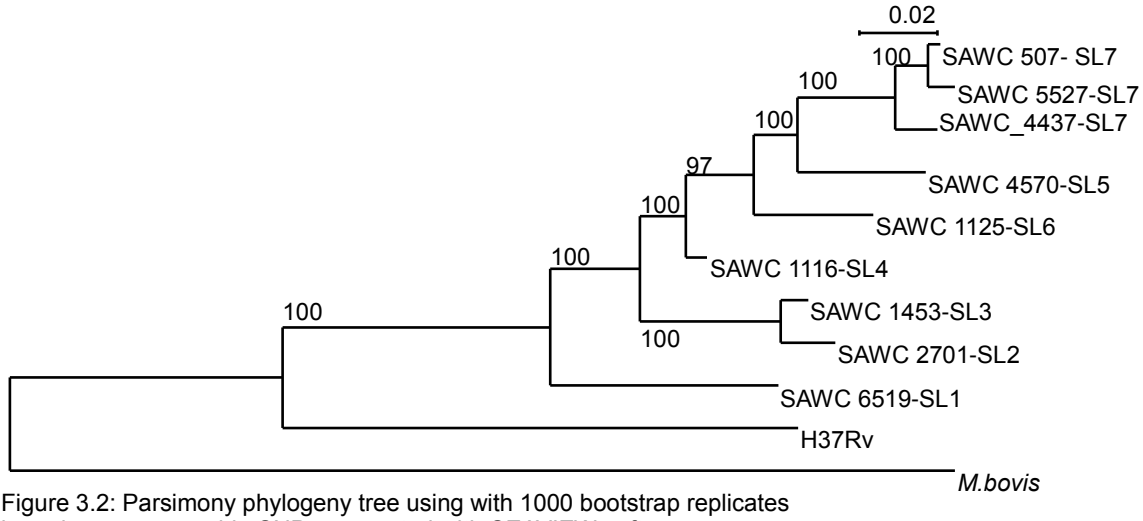


Figure 3.2: Parsimony phylogeny tree using with 1000 bootstrap replicates based on genome-wide SNPs generated with SEAVIEW software
*Sub-lineage (SL)

The topology of these trees was largely congruent with the evolutionary scenario predicted by Hanekom *et al*., 2007b with the exception of a reversal of order in which sub-lineages 5 and 6 occurred in the tree. In addition, isolates representing sub-lineages 2 and 3 were shown to be closely related and thereby did not reflect step-wise evolution as previously suggested by (Hanekom *et al*., 2007b).

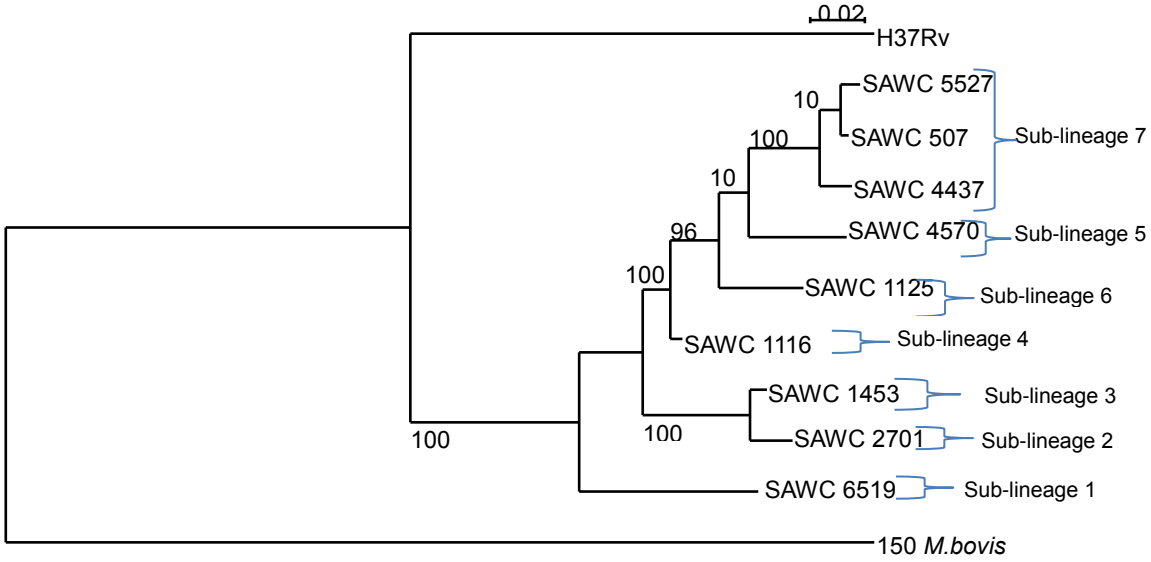


Figure 3.3: PHYML Maximum likelihood phylogeny tree using with 1000 bootstrap replicates based on genome-wide SNPs generated with SEAVIEW software.

3.8 Phylogeny based on 253 verified SNP positions

To determine whether the verified SNPs could accurately predict the phylogenetic history of the Beijing sub-lineages, trees were constructed using the data from the 273 SNP positions previously verified in a study by Schürch *et al*. (2011). Of the 273 verified positions, only 253 SNP positions were common among all 3 genome mapping software in our identification of high confidence SNPs in the 9 strains included in this analysis. Figure 3.4 shows the maximum likelihood tree generated from this set of SNPs which showed the same evolutionary scenario of strains as when using genome-wide SNPs in spite of the inclusion of strains from the study by Schürch *et al*. (2011). The topology of the tree constructed using only strains in this study using the 253 aforementioned SNPs was congruent with the tree based on the genome-wide SNP set with a patristic correlation value of 1.

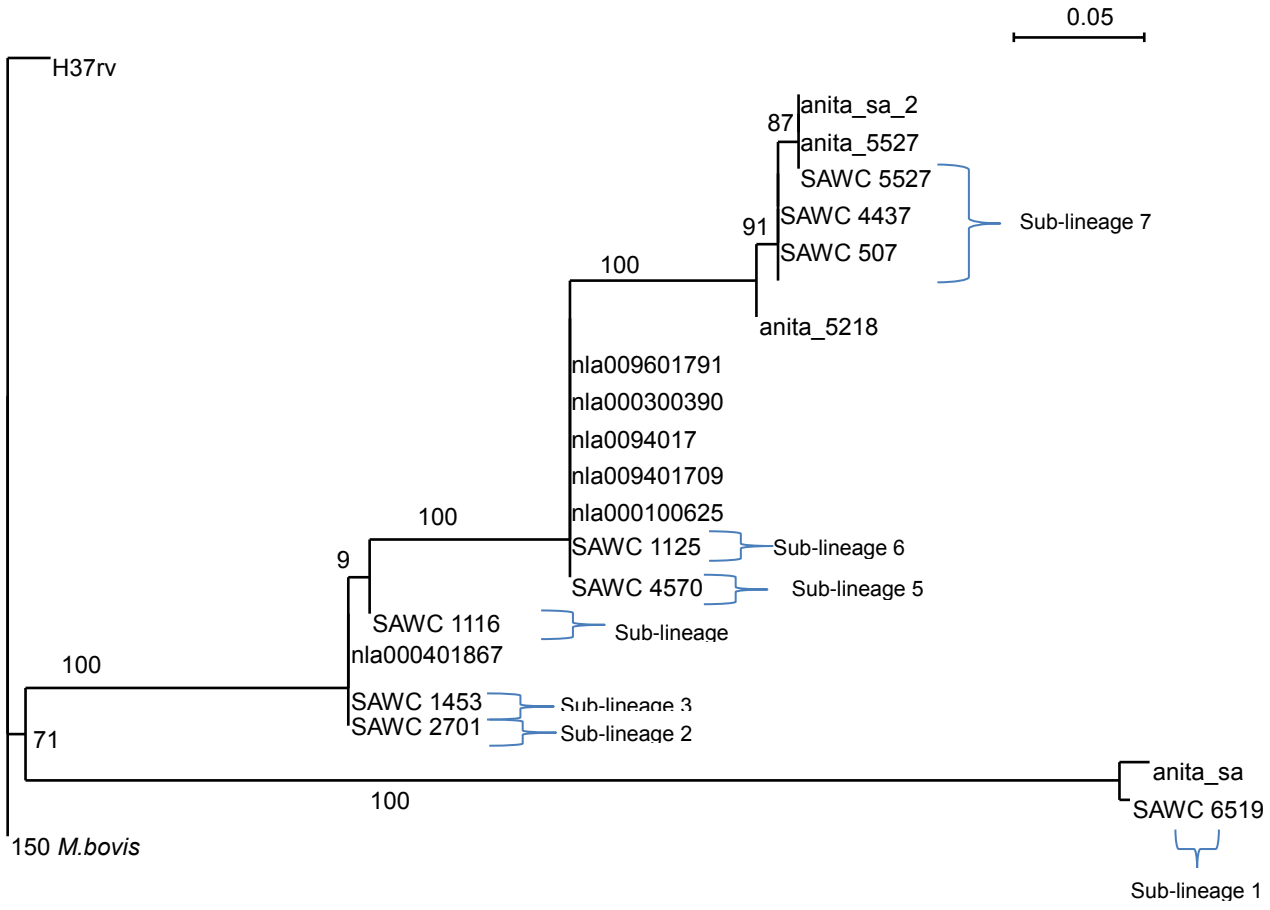


Figure 3.4: PHYML Maximum likelihood phylogeny tree using verified SNP positions of SAWC 507 and SAWC 5527 with 1000 bootstrap replicates based on genome-wide SNPs generated with SEAVIEW software. Strain s with the prefix “nla” are from the Netherlands.

Strains from a Dutch study by Schürch *et al*. (2011) were included in this phylogeny analysis in order to assess whether the phylogeny trees constructed in this study were influenced by selection bias as a result of not having a global representation of strains.

3.9 Phylogeny based on whole genome SNPs excluding non-synonymous SNPs

Synonymous SNPs have been previously used to identify lineage specific SNPs as they are considered to be evolutionarily neutral (Coll *et al*., 2014). The exclusion of nsSNPs resulted in the generation of trees with identical topology to those based on the genome-wide SNPs (Figure 3.3 and 3.5) with a patristic distance correlation of 1. However, the topology of the trees based on sSNPs was identical to the nsSNP tree with a patristic distance correlation of 1.

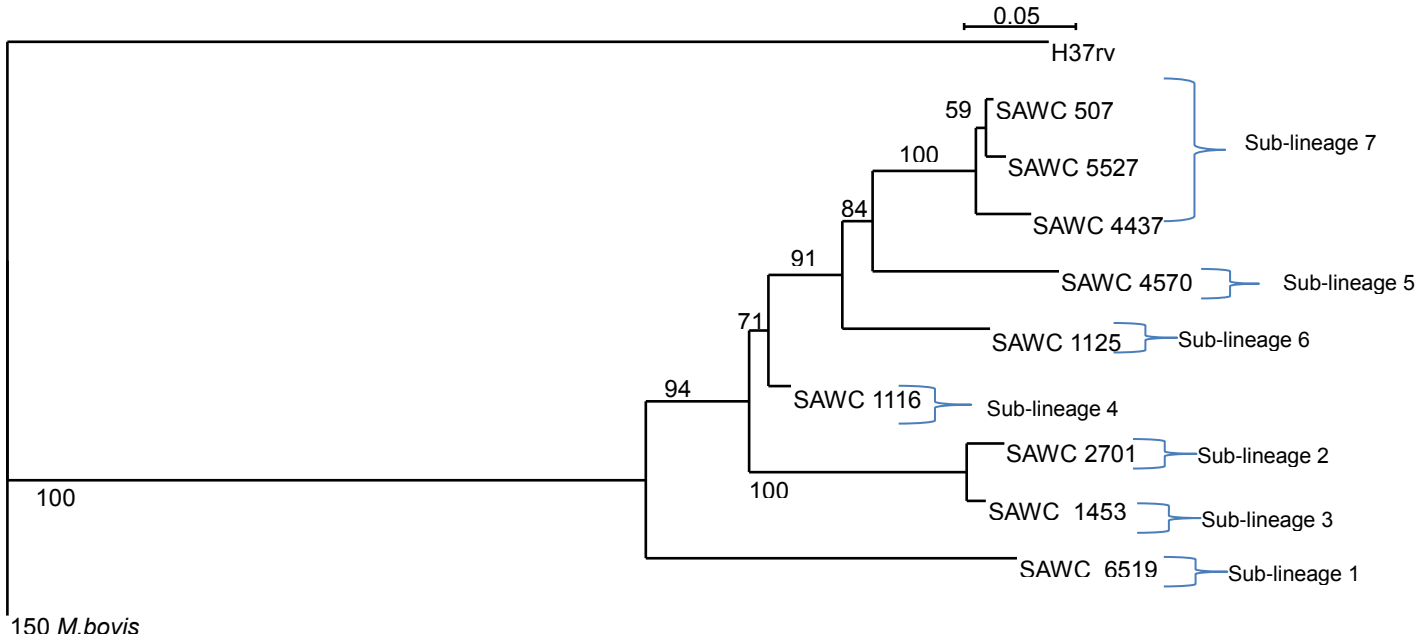


Figure 3.5: PhyML Maximum phylogeny tree using with 1000 bootstrap replicates based on genome-wide sSNPs generated with SEAVIEW software.

3.10 Phylogeny based on informative set of SNPs

The informative sSNPs for strains in this study was computed using SNPT. An analysis of phylogenetic trees constructed using informative sSNPs was compared to that involving genome-wide sSNPs. This revealed that the trees had the same topology as shown in Figures 3.5, 3.6 and 3.7 and had a patristic correlation coefficient of 1.

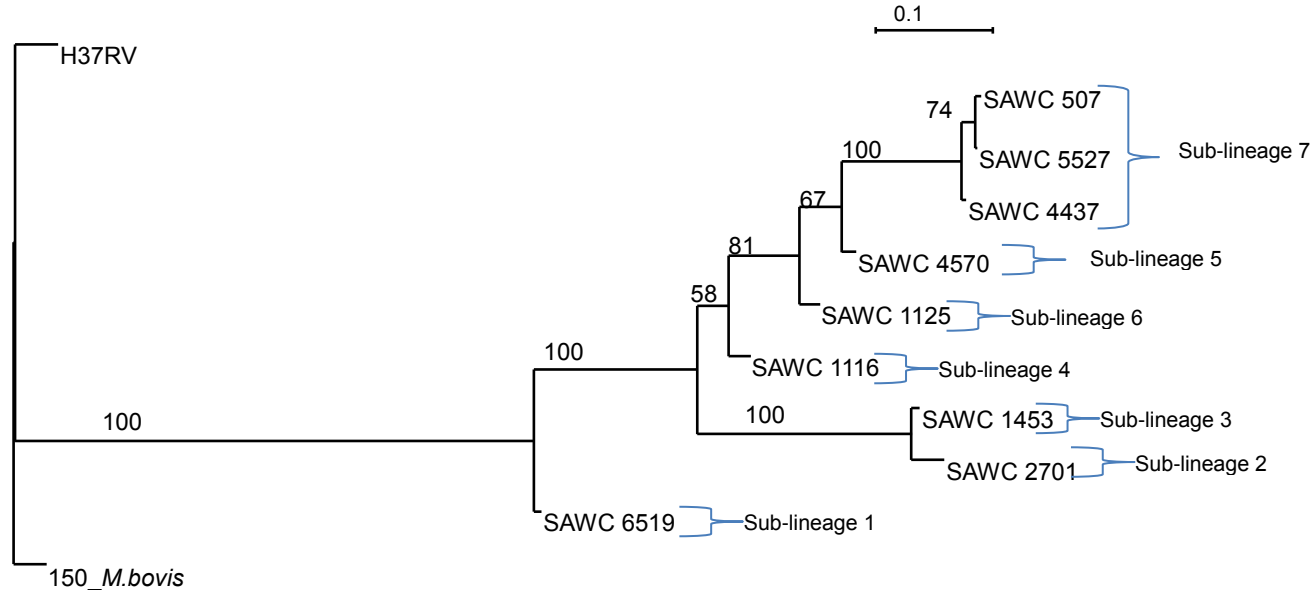


Figure 3.6: Parsimony phylogenetic tree with 1000 bootstrap replicates based on 288 informative SNPs generated with SEAVIEW software (bootstrap value indicted at nodes).

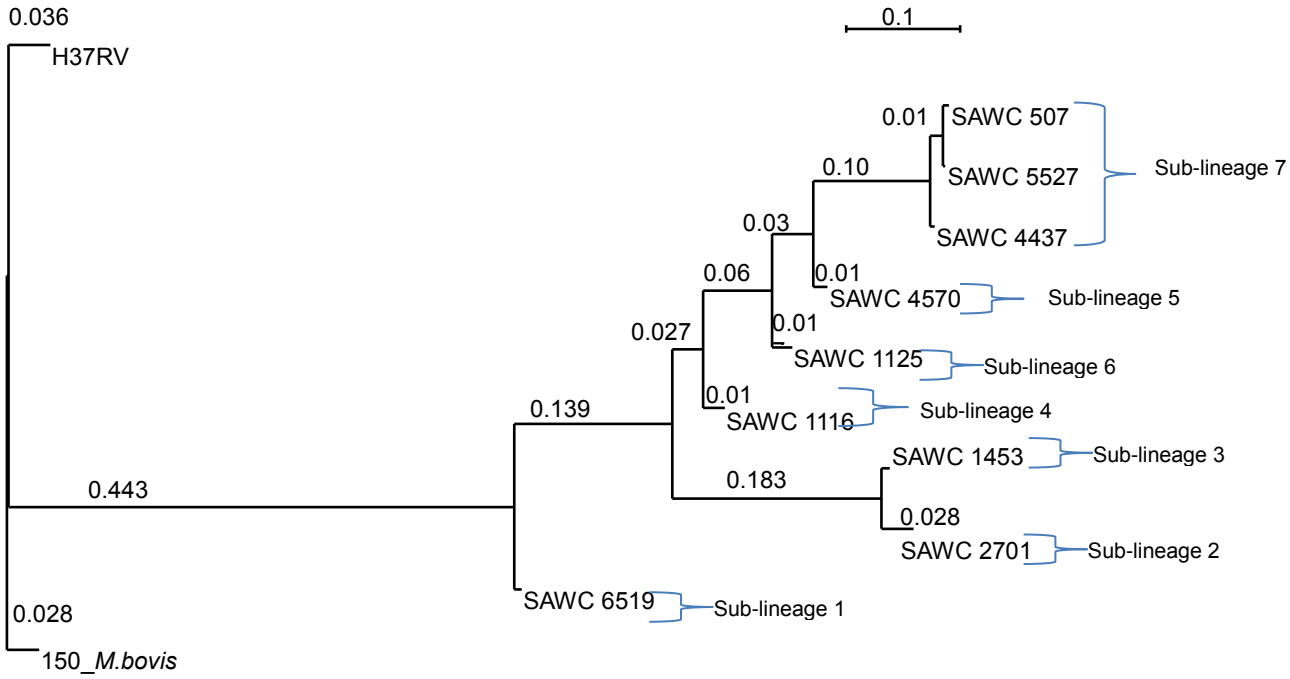


Figure 3.7: Parsimony phylogenetic tree with 1000 bootstrap replicates based on 288 informative SNPs generated with SEAVIEW software (branch lengths indicated at nodes).

3.11 Comparison of global phylogeny trees to replication, repair and recombination (3R) system phylogeny trees

The use of SNPs found in the genes of the Replication, Repair and Recombination (3R) system have previously been suggested to have high resolution in discriminating Beijing strains. To test this hypothesis we constructed trees based on SNPs present in the 3R genes and compared it to that using genome-wide SNPs. The topology of the 3R SNP trees were compared to the topology of the trees based on the global genome-wide sSNPs. The 3R SNP trees grouped the Beijing strains into only two braches (Figure 3.8.) demonstrating the inability of the 3R gene SNPs to accurately predict the evolutionary history of the Beijing lineage. The correlation between the phylogenetic tree based on selected 3R gene SNPs and that based on genome-wide sSNPs was a patristic distance correlation of 0.8034 where a patristic distance correlation of 1 indicates 100% correlation.

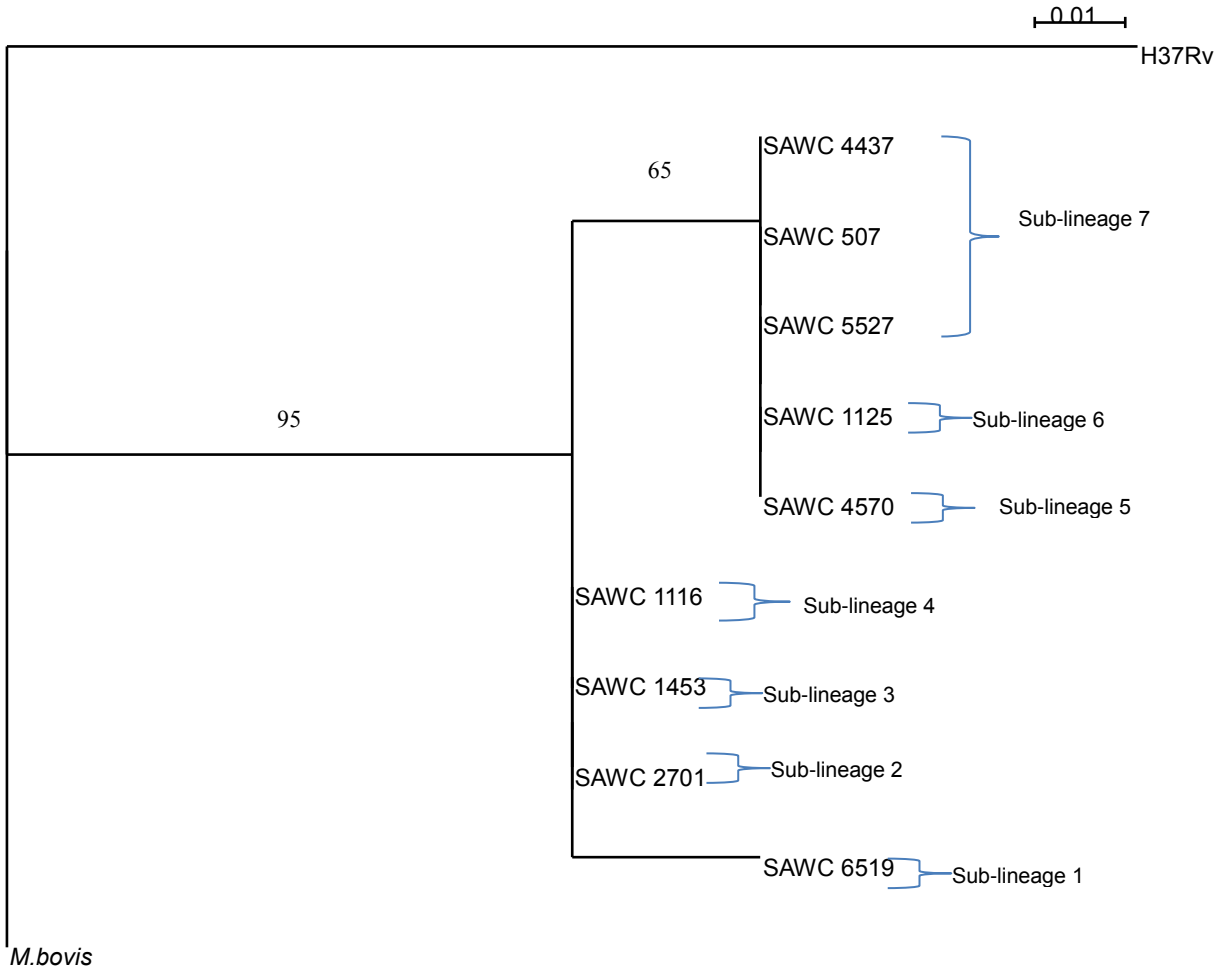


Figure 3.8: Parsimony phylogeny tree with 1000 bootstrap replicates generated with SEAVIEW software using 48 3R gene sites of which 4 were informative.

3.12 Comparative Analysis of Non-synonymous SNPs in the Evolution of 7 Sub-lineages of Beijing

The evolution of the Beijing lineage in this study is illustrated in Figure 3.9. We specifically sought to identify the nsSNPs that are unique to each of the sub-lineages that show how the lineages evolved with respect to amino acid changes of which the results are given in Table 3.2.

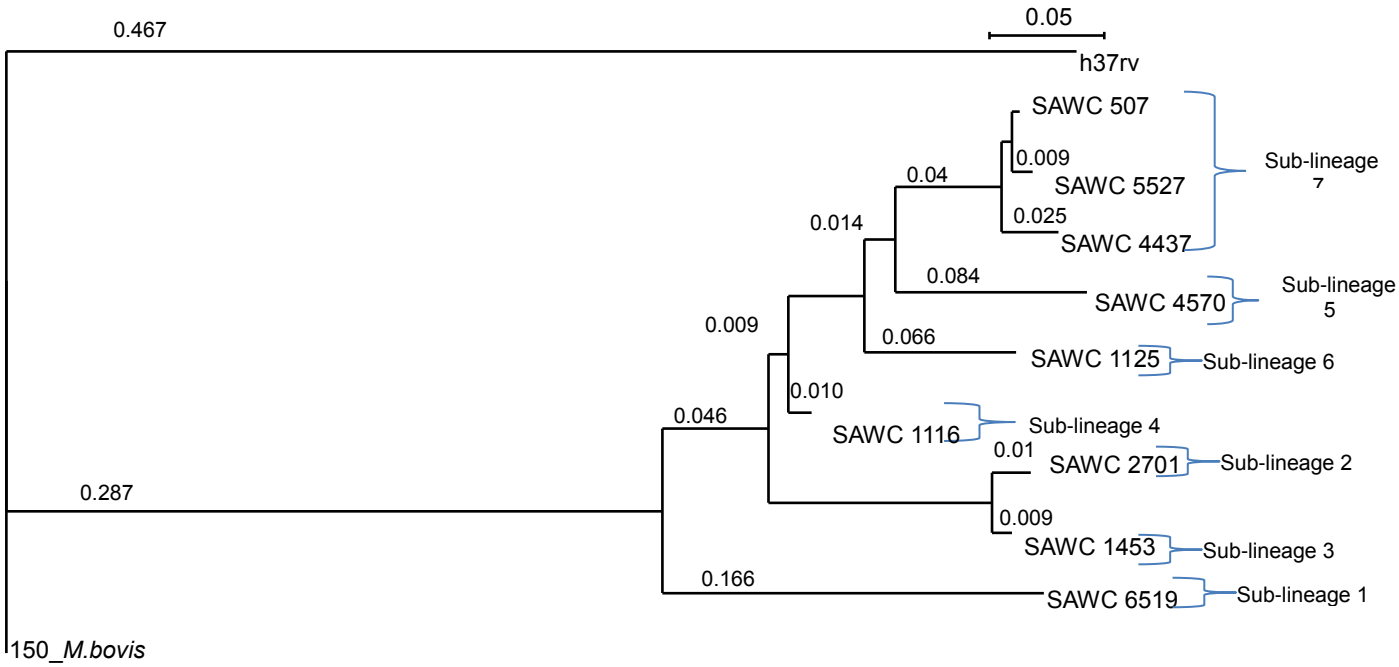


Figure 3.9: PhyML Maximum phylogeny tree showing branch lengths based on genome-wide sSNPs generated with SEAVIEW software

The unique nsSNPs were grouped into functional categories according to Tuberculist (Lew *et al*., 2011). In cases where 2 sub-lineages share a node on the phylogenetic tree and exhibited minor unique nsSNPs, the two sub-lineages were combined in the analysis of which the results are indicated by an asterisk (*). This was the case for sub-lineages 2 and 3.

Table 3.2: Functional distribution of SNPs unique to Beijing sub-lineages.

	Sub-Lineage 1		Sub-Lineage 2		Sub-Lineage 3		Sub-Lineage 4		Sub-Lineage 5		Sub-Lineage 6		Sub-Lineage 7	
Functional Group	**sSNPs**	**nsSNPs**	**sSNPs**	**nsSNPs**	**sSNPs**	**nsSNPs**	**sSNPs**	**nsSNPs**	**sSNPs**	**nsSNPs**	**sSNPs**	**nsSNPs**	**sSNPs**	**nsSNPs**
Information pathways	6	13	0 (6)*	0 (4)*	0 (6)*	2 (4)*	0	7	2	1	1	3	0	1
Lipid metabolism	7	13	0 (3)*	1 (12)*	0 (3)*	0 (12)*	0	9	4	10	6	9	4	7
Intermediary metabolism and respiration	29	30	1 10)*	0 (23)*	1 (10)*	2 (23)*	1	18	13	18	8	11	9	14
Cell wall and cell processes	16	37	1 (13)*	0 (30)*	1 (13)*	3 (30)*	1	17	13	8	10	12	6	16
Conserved hypotheticals Proteins	13	27	0 (10)*	0 (25)*	2 (10)*	0 (25)*	2	8	8	15	5	21	7	14
Virulence, detoxification, adaptation	1	7	0 (0)*	0 (4)*	0 (0)*	0 (4)*	0	2	0	4	2	3	2	2
PE/PPE	11	8	0 (6)*	0 (14)*	6 (6)*	0 (14)*	6	7	2	8	3	10	0	7
Insertion sequences and phages	3	2	0 (5)*	0 (1)*	5 (5)*	1 (1)*	5	3	2	1	2	0	1	0
Regulatory proteins	2	8	0 (3)*	0 (4)*	1 (3)*	0 (4)*	1	2	2	5	3	5	1	4
Intergenic SNPs	32		0 (20)*		0 (20)*		27		14		15		9	

* Node on the phylogenetic tree and exhibited minor unique nsSNPs and the two sub-lineages were combined in the analysis

From Figure 3.9 sub-lineage 1 appears to share a common ancestor with the rest of the sub-lineages after which it branched out separately. The majority of nsSNP unique to sub-lineage 1 were located in genes involved in intermediary metabolism and respiration, cell wall and cell processes and conserved hypothetical functional categories as highlighted in Figure 3.10.

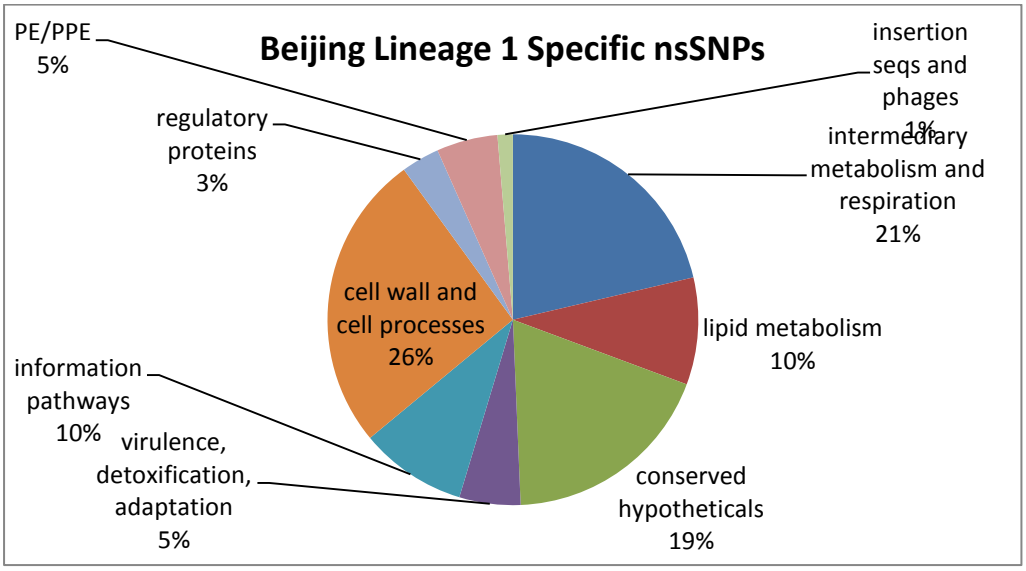


Figure 3.10: Functional distribution of genes containing SNPs unique to Beijing Sub-lineage 1.

Sub-lineages 2 and 3 share a common ancestor and the genetic distance between these lineages appears to be less than previously thought. This is depicted in Figure 3.9 where their branch distances from a common ancestor are 0.01 and 0.009 respectively in a maximum likelihood tree. The majority of unique nsSNPs were in the intermediary metabolism and respiration, cell wall and cell processes and conserved hypothetical functional categories as shown in Figure 3.11

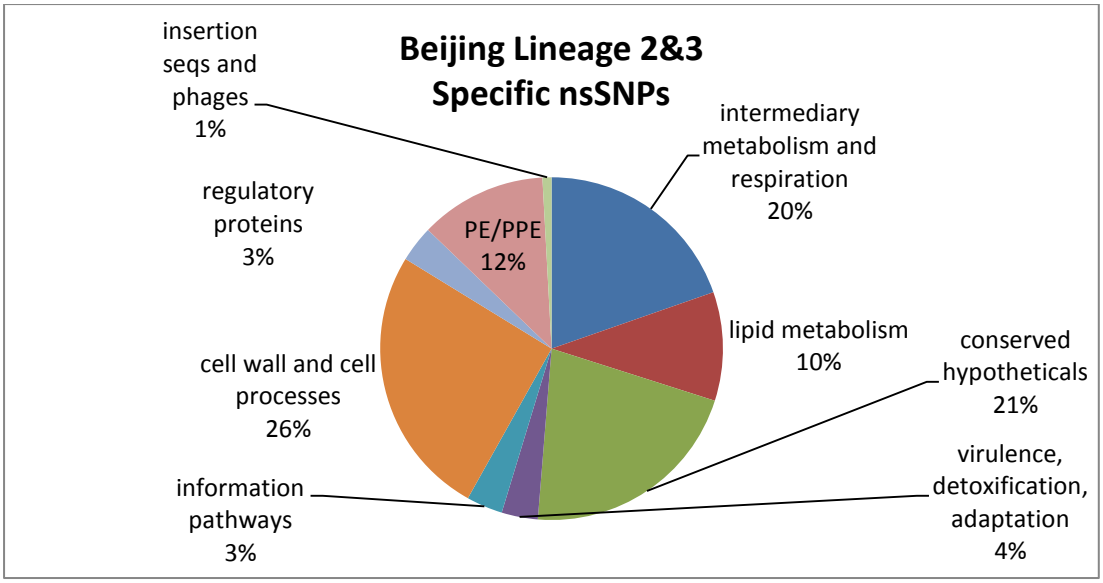


Figure 3.11: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 2&3.

The nsSNPs distribution by functional category for Beijing sub-lineage 4 in this study are illustrated in Figure 3.12 and showed that the majority unique nsSNPs to sub-lineage 4 are found in the intermediary metabolism and respiration, cell wall and cell processes functional categories. This is followed by nsSNPs in the conserved hypothetical and lipid metabolism functional categories.

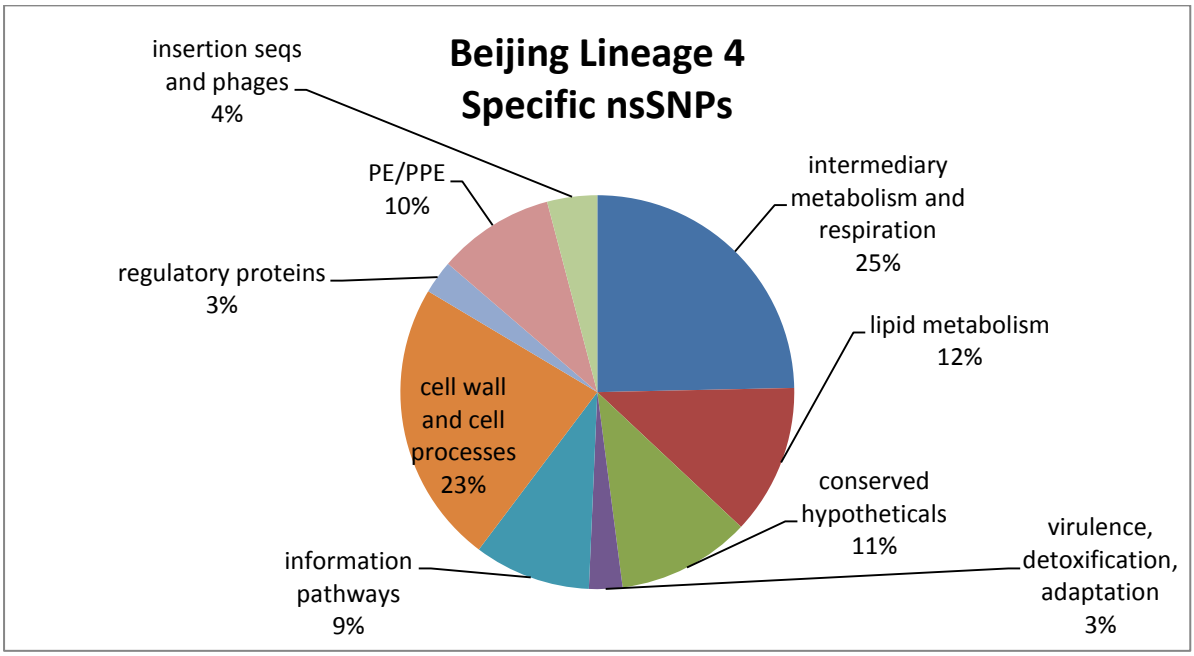


Figure 3.12: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 4.

The nsSNP functional categories in this study for Beijing sub-lineage 5 are illustrated in Figure 3.13. Intermediary metabolism and respiration and conserved hypothetical functional categories accounted for the bulk of nsSNPs. This was followed by lipid metabolism, PE/PPE and cell wall and cell processes. Insertion sequences and phages and information pathway categories had the least representation.

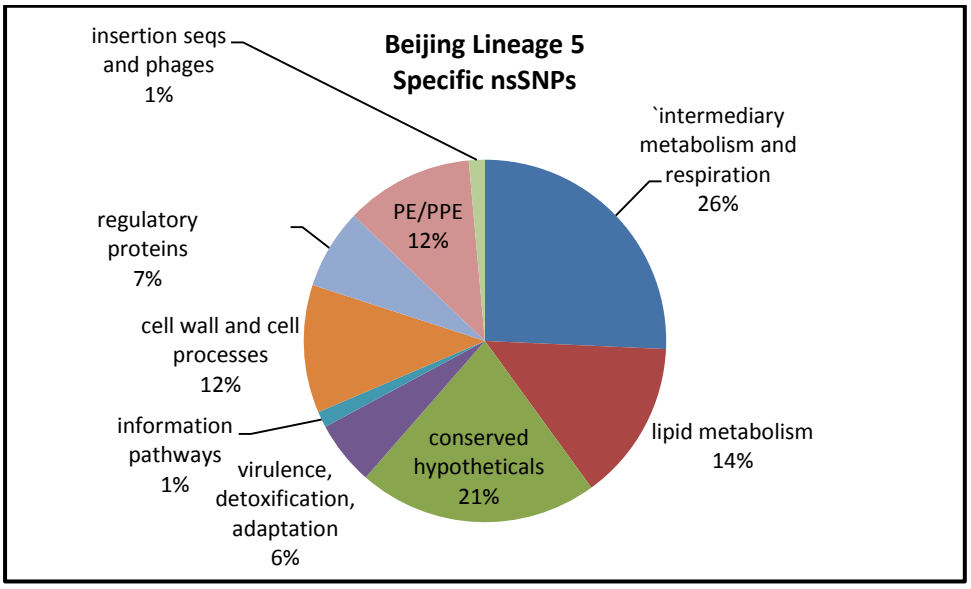


Figure 3.13: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 5.

The functional distribution of nsSNPs unique to Beijing sub-lineage 6 in this study is shown in Figure 3.14. In the present study, the conserved hypothetical functional group comprised the majority of nsSNps unique to Beijing sub-lineage 6. This was followed by nsSNPs in the intermediary metabolism and respiration, cell wall and cell processes and lipid metabolism functional groups.

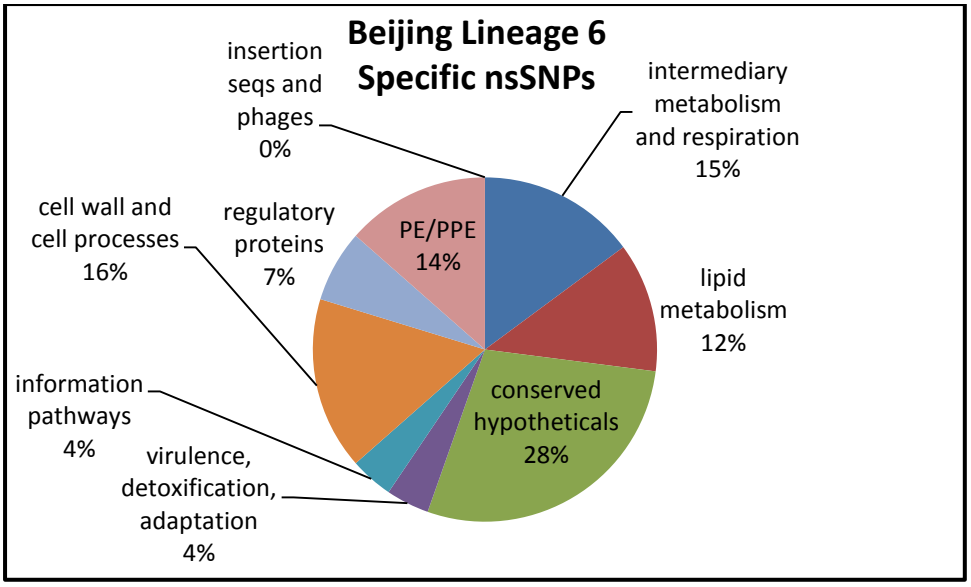


Figure 3.14: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 6.

In this study the nsSNPs unique to Beijing sub-lineage 7 are shown in Figure 3.15 and show that cell wall and cell processes, conserved hypotheticals and intermediary metabolism and respiration functional accounted for the majority of unique nsSNPs.

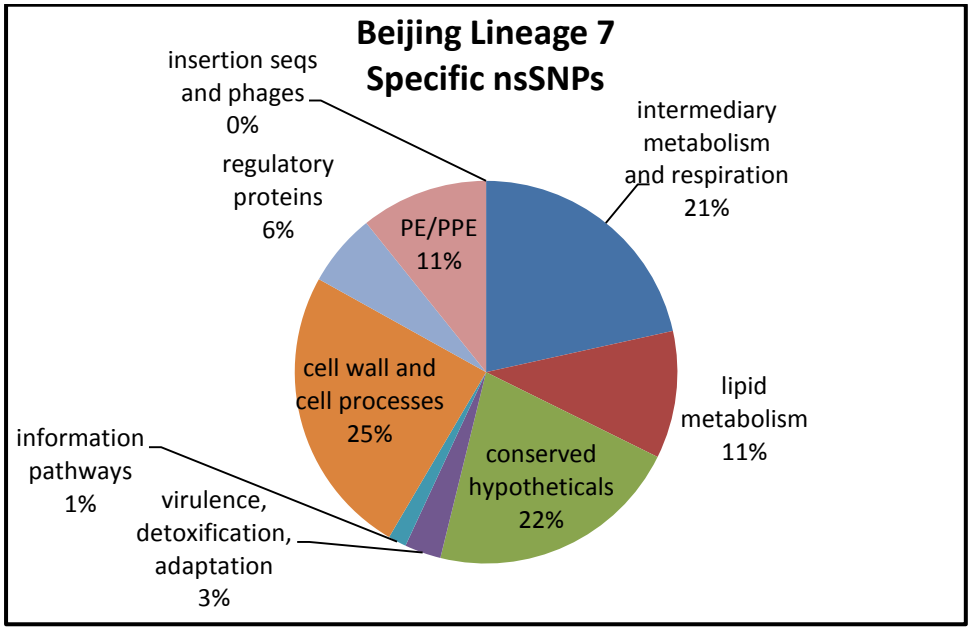


Figure 3.15: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 7.

3.13 Non-Synonymous SNPs common to each node (Branch Point)

In order to further elucidate the evolutionary scenario of Beijing strains include in this study, an analysis was undertaken to determine the functional relevance of sequence changes at branch points of the evolutionary trees. The results of the branch point or node analysis are given in Table 4.3. The analysis showed how many SNPs were common to all Beijing strains including what was accumulated at each particular node as is summarised in Figure 3.16.

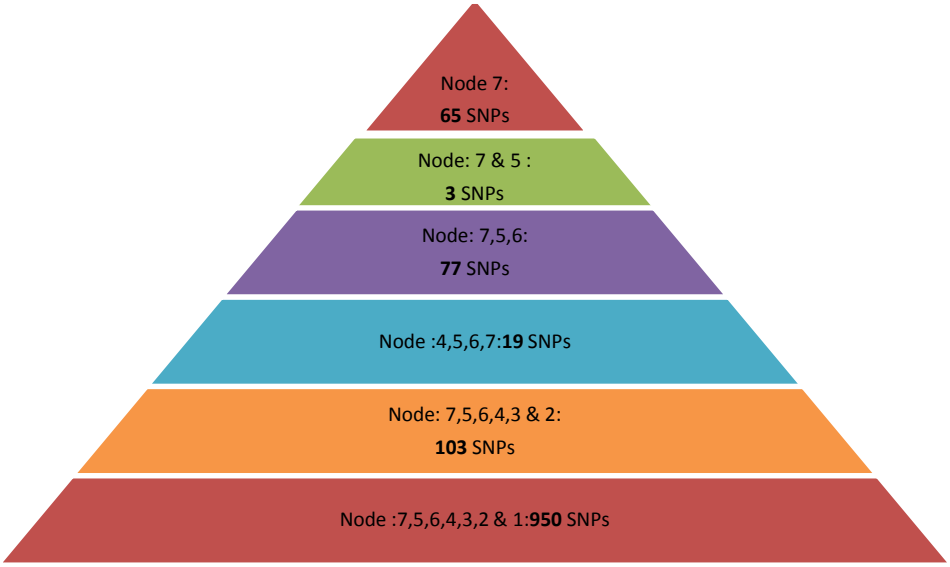


Figure 3.16: Number of functional distribution of SNPs at the nodes of phylogenetic tree where the nodes comprise the common ancestor to sub-lineages specified in the node name. Node 7,5,6 in this instance means that node 7,5,6 represents the common ancestor of sub-lineages 7,5, and 6 as reflected in the genome-wide phylogenetic trees and has 77 SNPs common to it.

Table 3.3: Functional distribution of SNPs at phylogenetic tree nodes (Branch Points) of the respective Beijing sub-lineages.

	Node 1= Sub-lineage 1, 2, 3, 4, 5, 6, 7			**Node 2= Sub-lineage 2, 3, 4, 5, 6, 7**			**Node 3= Sub-lineage 4, 5, 6, 7**			**Node 4= Sub-lineage 5, 6, 7**			**Node 5= Sub-lineage 5, 7**			**Node 6= Sub-lineage 7**		
Functional Group	sSNPs	nsSNPs	Total	sSNPs	nsSNPs	Total	sSNPs	nsSNPs	Total	sSNPs	nsSNPs	Total	sSNPs	nsSNPs	Total	sSNPs	nsSNPs	Total
Information Pathways	21	39	60	1	6	7	0	1	1	1	3	4	0	0	0	0	1	1
Lipid Metabolism	43	55	98	11	5	16	0	3	3	1	3	4	0	0	0	4	7	11
Intermediary Metabolism and Respiration	64	128	192	12	20	32	2	2	4	3	11	14	0	0	0	9	14	23
Cell Wall and Cell Processes	83	135	218	7	5	12	0	2	2	6	10	16	0	0	0	6	16	22
Conserved Hypotheticals	66	114	180	5	9	14	1	2	3	6	9	15	0	0	0	7	14	21
Virulence, Detoxification, Adaptation	14	17	31	2	2	4	0	2	2	1	5	6	0	0	0	2	2	4
Regulatory Proteins	19	20	39	1	4	5	0	1	1	0	7	7	0	0	0	0	4	4
PE/PPE	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	1	7	8
Insertion Sequences and Phage	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Intergenic SNPs	132			13			3			11			1			9		
Grand Total SNPs	950			103			19			77			3			104		

* Node on the phylogenetic tree and exhibited minor unique nsSNPs and the two sub-lineages were combined in the analysis

The common ancestor node of sub-lineages 1, 2, 3, 4, 5, 6and 7 had the majority of nsSNPs in the Tuberculist functional groupings of intermediary metabolism and respiration, cell wall and cell processes and conserved hypotheticals groups as shown in Figure 3.17. No nsSNPs were identified in the PE/PPE and insertion sequences and phages functional groupings.

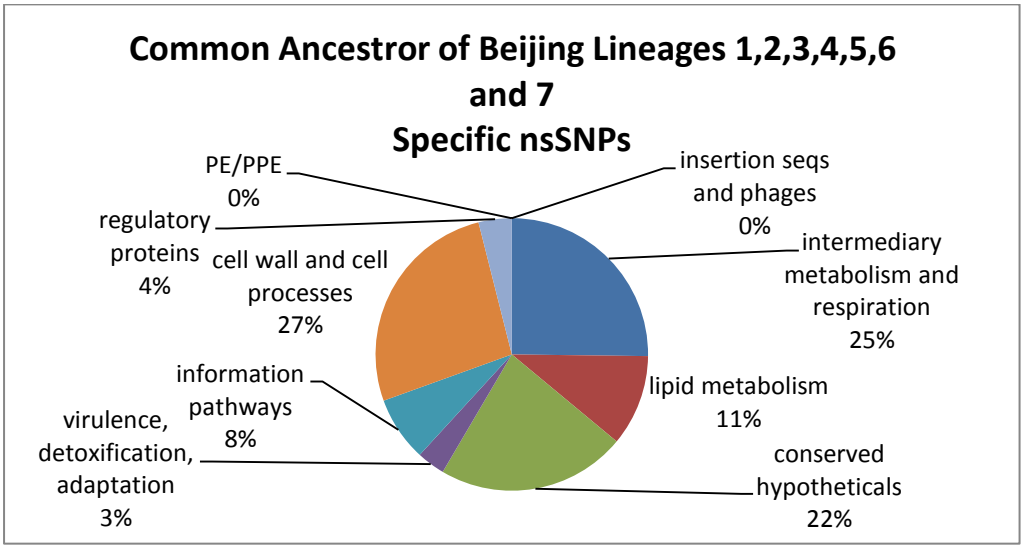


Figure 3.17: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 1, 2, 3, 4, 5, 6 and 7.

The majority of nsSNPs in the common ancestor of sub-lineages 2, 3, 4, 5, 6 and 7 were in the intermediary metabolism and respiration Tuberculist functional groupings and illustrated in Figure 3.18. This was followed by nsSNPs in the groupings of conserved hypothetical proteins, information pathways, cell wall and cell wall processes and lipid metabolism. No nsSNPs were identified in the PE/PPE and insertion sequences and phages functional groupings

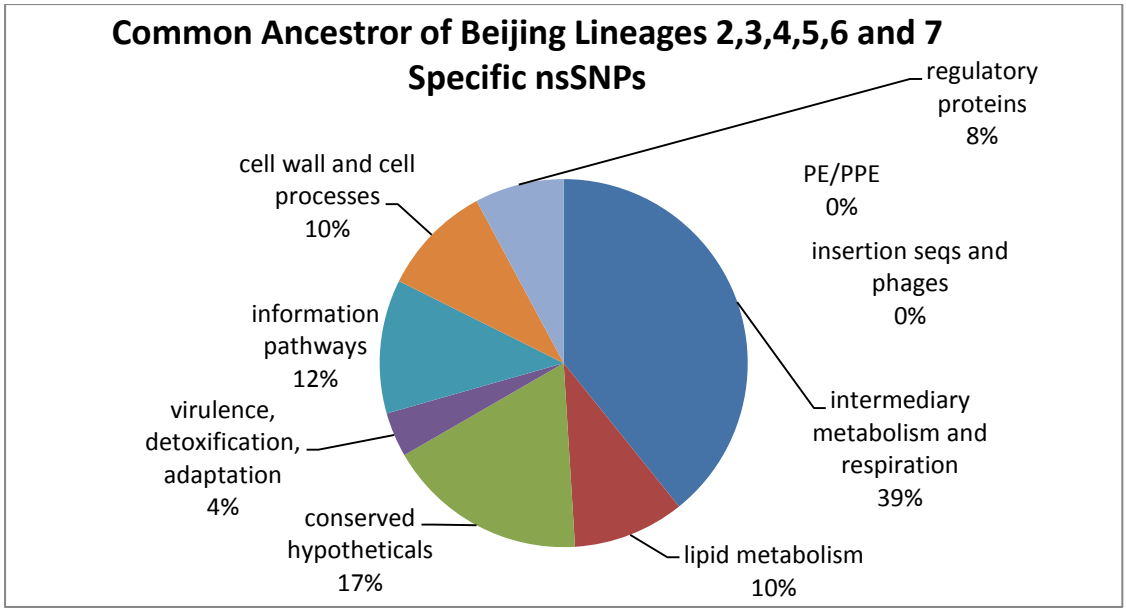


Figure 3.18: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 2, 3, 4, 5, 6 and 7

The common ancestor of sub-lineages 4, 5, 6, and 7 had a total of 13 nsSNPs common to it. The majority nsSNPs were in the lipid metabolism grouping of the Tuberculist functional groupings followed by generally equal distribution among of intermediary metabolism and respiration, cell wall and cell processes, conserved hypotheticals groups. The PE/PPE and insertion sequences and phages functional grouping which had no SNPs found in them as illustrated in Figure 3.19.

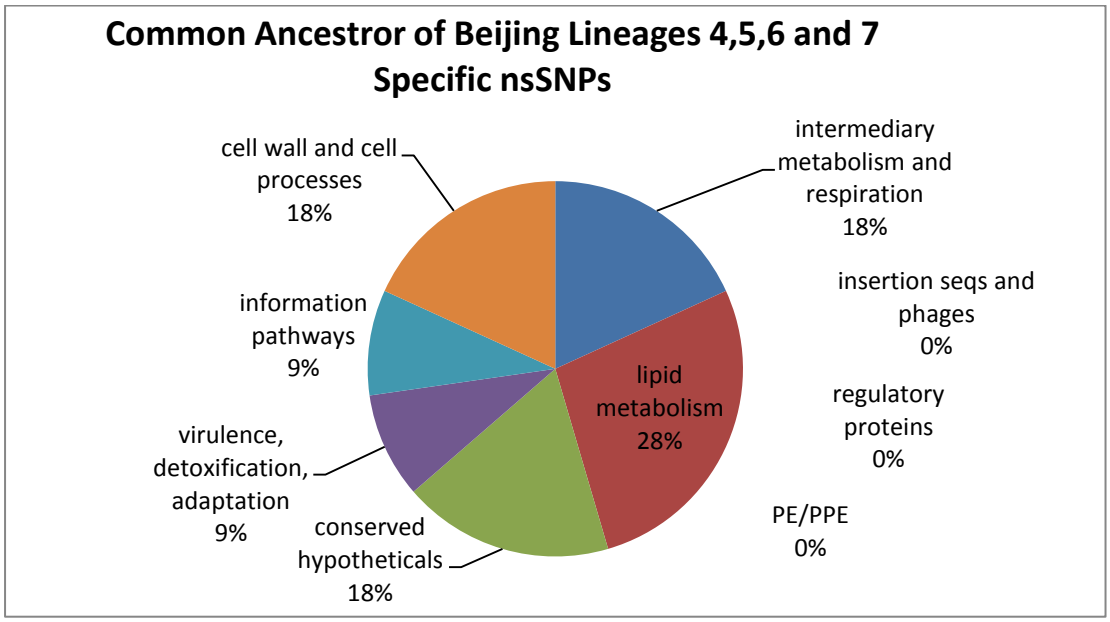


Figure 3.19: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 4, 5, 6 and 7

There were 77 nsSNPs which were unique to the common ancestor for the node encompassing sub-lineages 5, 6 and 7. The majority of these SNPs were in the functional categories of intermediary metabolism and respiration, cell wall and cell processes, conserved hypotheticals and regulatory functional groupings as depicted in Figure 3.20.

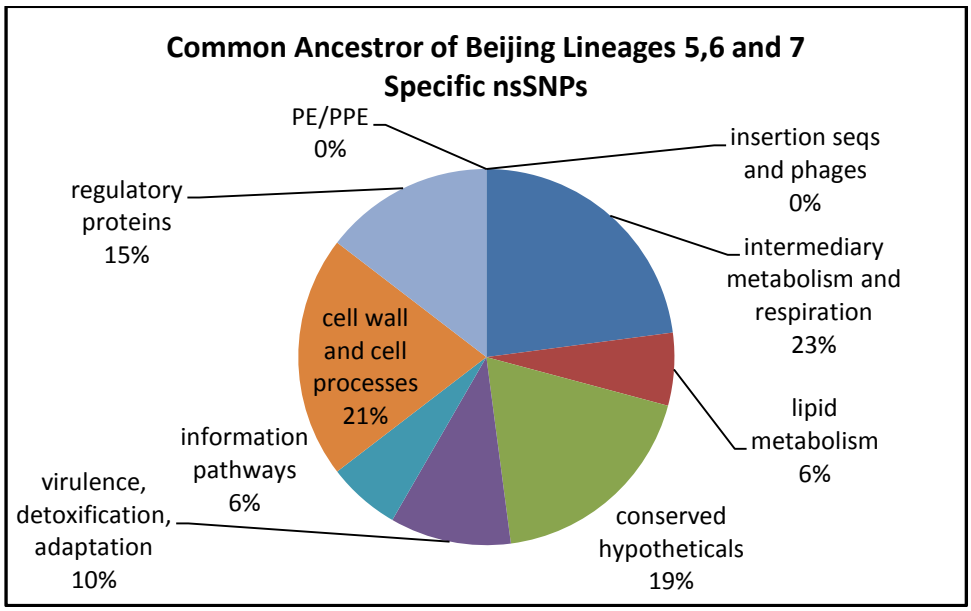


Figure 3.20: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 5,6 and 7

The common ancestor node of sub-lineages 5 and 7 only had 1 nsSNP in the PE/PPE functional groupings.

3.14 Biological processes functional evolution of the 7 sub-lineages of Beijing

The identification of overrepresented gene ontology (GO) terms in genes containing nsSNPs that are unique to sub-lineages contributes to our understanding of the evolution of Beijing sub-lineages in this study. Use of whole genome nsSNP data and the PANTHER (protein annotation through evolutionary relationship) (Mi *et al*., 2013) classification system resulted in the identification of overrepresented GO terms in the biological processes domain that were unique to sub-lineages or common ancestors as depicted in the whole genome phylogenetic trees constructed in this study (Figure 3.9). The analysis showed that there was overrepresentation of GO terms involved in biological processes whose genes had unique sub-lineage nsSNP only in sub-lineages 1 and combined sub-lineages 2 and 3. The biological processes that were overrepresented for sub-lineage 1 in this regard were the primary metabolic processes and metabolic process as given in Table 3.4. In the case of sub-lineage 2 and 3, protein acetylation, cellular protein modification processes and protein metabolic processes were overrepresented as shown in Table 3.5.

Table:3.4: PANTHER overrepresentation of biological processes whose genes had unique sub-lineage nsSNP only in sub-lineage 1.

PANTHER Process GO-Slim Biological Process	Fold Enrichment	Fold Enrichment +/-	P value
Primary metabolic process	1.65	+	6.06E-03
Metabolic process	1.64	+	1.61E-04
Unclassified	0.82	-	0.00E+00

Table:3.5: PANTHER overrepresentation of biological processes whose genes had unique sub-lineage nsSNP only in sub-lineages 2 and 3.

PANTHER GO-Slim Biological Process	expected	Fold Enrichment	Fold Enrichment +/-	P value
Protein acetylation	0.58	> 5	+	1.97E-04
Cellular protein modification process	2.31	> 5	+	3.46E-04
Protein metabolic process	6.53	2.91	+	2.33E-03
Unclassified	77.9	0.87	-	0.00E+00

Focusing on nodes of phylogenetic trees which represent common ancestors of branches that originate from a node, only the nsSNPs defining the node encompassing sub-lineages 2, 3, 4, 5, 6 and 7 had an overrepresentation of GO terms involved in biological processes. This was in the category of protein metabolic processes as shown in Table 3.6.

Table 3.6: PANTHER overrepresentation of biological processes whose genes had sub-lineage nsSNP only in sub-lineages 2,3,4,5,6 and 7 branch point common ancestor.

PANTHER GO-Slim Biological Process	Fold Enrichment	Fold Enrichment +/-	P value
Metabolic process	1.85	+	2.66E-02
Unclassified	0.79	-	0.00E+00

3.15 Comparative analysis of transcriptional start sites and promoters in the evolution of 7 sub-lineages of Beijing

All intergenic or non-coding (nc) SNPs were initially analysed irrespective of their downstream distance from the gene start sites. A cut off of 500bp was then used based on the TSS distance from the start of a gene for identified ncSNPs as depicted in Figure 3.21 in this study.

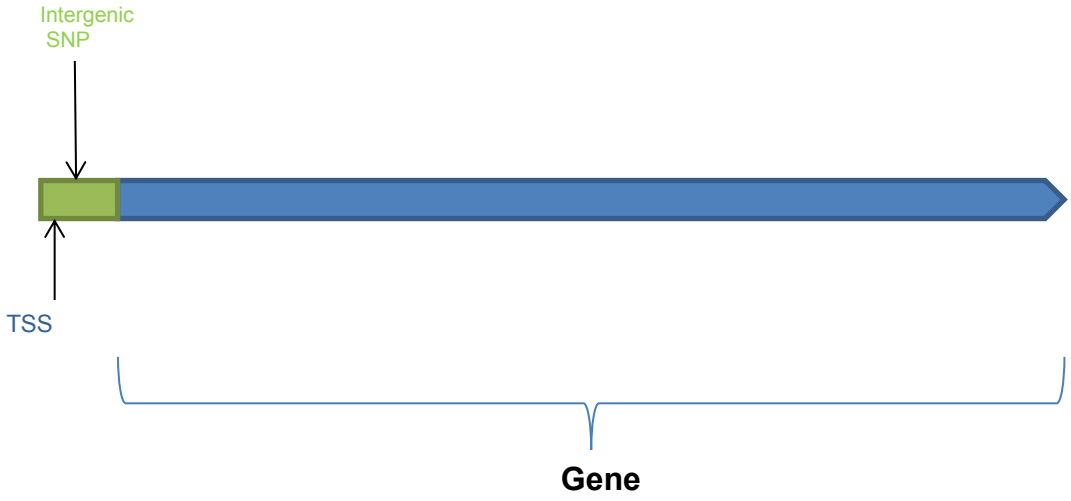


Figure 3.21: Functional distribution of genes containing SNPs common to Beijing Sub-lineage 2&3

The TSSs are only reported for those that have been identified in the *M. tuberculosis* H37Rv reference genome, as per study by Cortes *et al.* (2013). Any non-coding SNPs which were not before the start of the nearest genes with respect to the orientation of a gene had a negative distance value from the start site of a gene. These are highlighted red in Tables 3.7, 3.8, 3.9, 3.10 and 3.11.

The Beijing sub-lineage 1 had 33 unique ncSNPs. Of these, 9 SNPs were not before the start site of the nearest gene or within 500bp of the start of a gene. These are depicted in red highlight in Table 3.7. From the 29 SNPs which were present before the start site of nearest gene, 16 genes had TSSs which have been reported in *M. tuberculosis* H37Rv by Cortes *et al*. (2013). Transcription start sites more than 35bp from the start codon of a gene were also observed for sub-lineage 1 in this study. These included *Rv0755c* whose transcriptional start site has been reported to be 466bp from the gene start site. Only *Rv0317c* and *Rv2347c* had TSSs within the 35bp region from the start of the gene.

Table 3.7: Sub-lineage 1- distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS)

Nearest Gene	Orientation	Sub-Lineage 1 Intergenic SNP	Distance from Gene	TSS	Gene Start Position
Rv0022c	-	27455	13	#N/A	27442
Rv0316	+	385158	-623	#N/A	384535
Rv0644c	-	739262	102	739243	739160
Rv0755c	-	850225	185	850506	850040
Rv0916c	-	1021776	133	1021685	1021643
Rv1067c	-	1190472	48	#N/A	1190424
Rv1128c	-	1253073	101	#N/A	1252972
Rv1147	+	1276232	-1332	#N/A	1274900
Rv1149	+	1276232	1661	1277893	1277893
Rv1430	+	1606248	138	1606253	1606386
Rv1483	+	1673425	15	1673440	1673440
Rv1530	+	1731293	80	#N/A	1731373
Rv1946c	-	2198515	555	#N/A	2197960
Rv1964	+	2207253	447	2207614	2207700
Rv1989c	-	2232577	-722	#N/A	2233299
Rv2329c	-	2603515	54	2603499	2603461
Rv2347c	-	2626600	81	2626552	2626519
Rv2769c	-	3079190	205	#N/A	3078985
Rv2842c	-	3150049	73	3150067	3149976
Rv2876	+	3187987	-324	#N/A	3187663
Rv3049c	-	3411154	71	3411169	3411083
Rv3122	+	3487716	373	#N/A	3488089
Rv3144c	-	3511368	51	3511427	3511317
Rv3179	+	3547142	476	#N/A	3547618
Rv3269	+	3650189	45	3650188	3650234
Rv3383c	-	3798732	243	#N/A	3798489
Rv3477	+	3894015	78	3893953	3894093
Rv3483c	-	3902910	98	#N/A	3902812
Rv3596c	-	4040976	272	4040759	4040704
Rv3823c	-	4291611	82	#N/A	4291529
Rv3863	-	4338794	55	4338714	4338849
Rv3901c	-	4386896	82	#N/A	4386365
Rv3920c	+	4408899	2	#N/A	4408906

SNP positions highlighted in red were not considered for further analysis because they were not within 500bp of the start of a gene

The total number of ncSNPs that met the criteria of being within 500bp downstream of a gene start site for sub-lineage 1 was 21. Eight of these had previously described TSS downstream of a gene start site. The relative distances of the ncSNPs from a

gene start site and associated TSS are depicted in Figure 3.22.

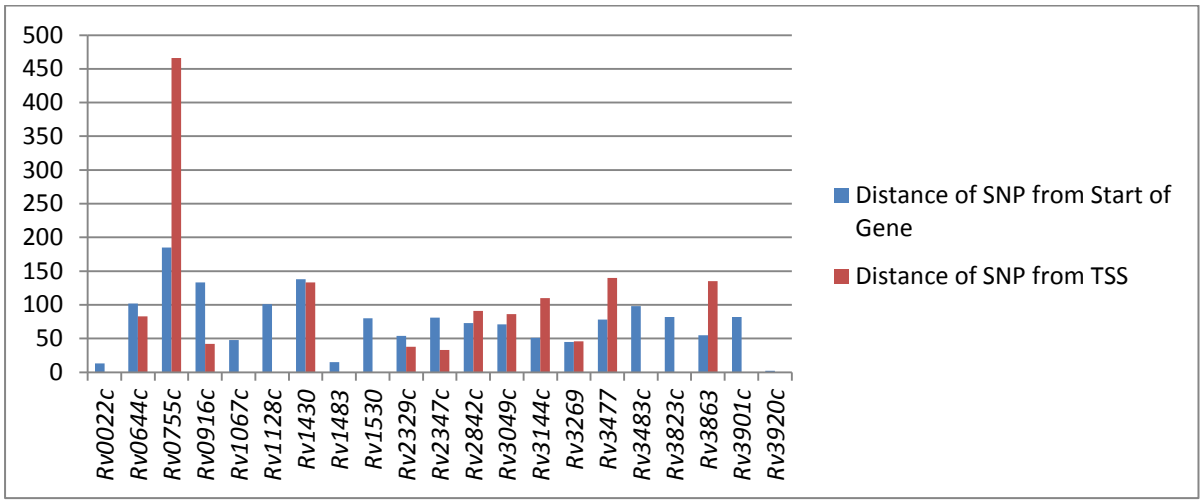


Figure 3.22: Graph illustrating the distance of ncSNP from the start of adjacent gene and from the genes reported transcription start site (TSS).

Sub-lineage 2 and 3 had a total of 21 unique ncSNP of which 6 were not situated before or within 500bp from the start of the nearest gene. Of the remaining 15 ncSNPs situated within 500bp of the start of the gene, 11 had TSS positions attributed to them as shown in Table 3.8.

Table 3.8: Sub-lineage 2 and 3 - distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

Nearest Gene	**Sub-Lineage** -2&3 **Intergenic** _SNP	Distance from Gene	TSS	Gene Start Position
Rv0103c	122189	16	122199	122173
Rv0750	842284	-251	842029	842033
Rv1093	1220548	26	1220538	1220574
Rv1147	1276232	-1332	#N/A	1274900
Rv1148c	1275923	-1825	#N/A	1277748
Rv1302	1459550	-1255	#N/A	1458295
Rv1303	1459550	216	1459707	1459766
Rv1861	2108954	211	#N/A	2109165
Rv1984c	2228790	229	2228561	2228561
Rv2331A	2604686	54	2604318	2604740
Rv2648	2972093	67	#N/A	2972160
Rv2785c	3093853	105	3093825	3093748
Rv2975c	3331905	293	3331657	3331612
Rv3129	3494574	86	#N/A	3494660
Rv3416	3834836	56	3834791	3834892

Rv3616c	4056972	597	4056443	4056375
Rv3617	4057190	543	4057733	4057733
Rv3646c	4087420	163	4087380	4087257
Rv3706c	4149940	29	#N/A	4149911
Rv3804c	4266885	227	4266750	4266658
Rv3888c	4372774	68	4372706	4372706

*SNP positions highlighted in red were not considered for further analysis because they were not within 500bp of the start of a gene

Sub-lineage 2 and 3 had a total of 10 ncSNPs within 500bp downstream of a gene. Three of the 10 ncSNPs had an associated TSS downstream of a gene start site. The relative distances of the ncSNPs to the TSS and gene start sites are illustrated in Figure 3.23.

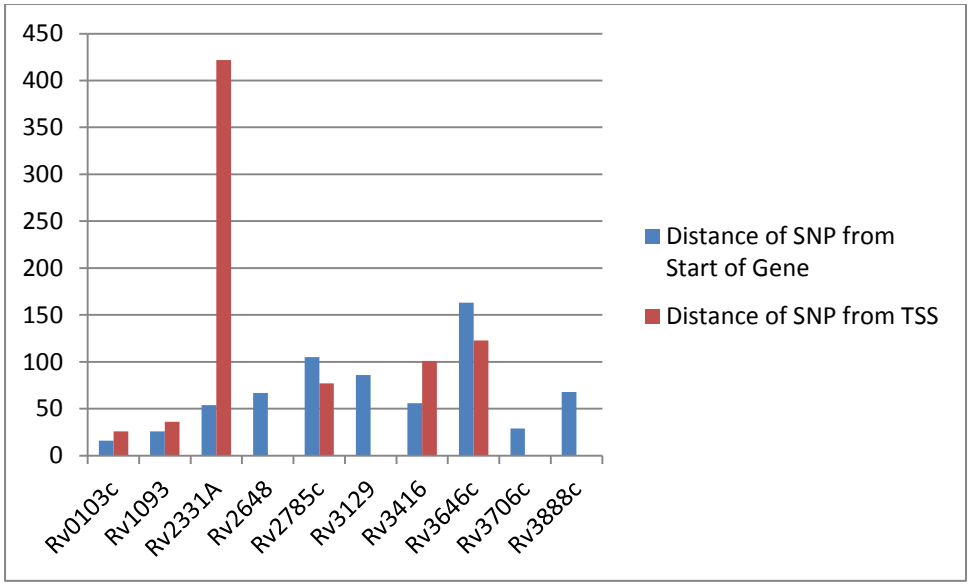


Figure 3.23: Graph illustrating the distance of ncSNP from the start of adjacent gene and from the genes reported transcription start site (TSS).

A total of 27 ncSNPs were identified for Beijing sub-lineage 4 in this study of which 9 ncSNPs were not within 500bp downstream of a gene start site as is shown in Table 3.9. A total of 14 genes had a TSS attributed to them.

Table 3.9: Sub-lineage 1distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

Nearest Gene	Orientation	*SL-4_IG_SNP	Distance from Gene	TSS	Gene Start Position
Rv0150c	-	177311	359	177341	176952
Rv0239	+	289067	37	289087	289104
Rv0458	+	549643	32	549644	549675
Rv0482	+	569841	698	#N/A	570539
Rv0488	+	577510	154	#N/A	577664
Rv0609A	+	704533	-703	#N/A	703830
Rv1040c	-	1164571	1195	#N/A	1163376
Rv1312	+	1468150	-462	#N/A	1467688
Rv1322	+	1485309	-327	#N/A	1484982
Rv1327c	-	1494456	31	1494480	1494425
Rv1356c	-	1524853	33	#N/A	1524820
Rv1375	+	1547697	135	1547599	1547832
Rv1443c	-	1622727	35	#N/A	1622692
Rv1511	+	1702632	442	1703020	1703074
Rv1607	+	1806027	154	1805922	1806181
Rv1719	+	1946456	-815	#N/A	1945641
Rv2005c	-	2251900	17	2251902	2251883
Rv2603c	-	2931622	62	2931593	2931560
Rv2815c	-	3123090	1263	#N/A	3121827
Rv2943	+	3287857	607	3287968	3288464
Rv2957	-	3309460	10	#N/A	3309470
Rv3190A	+	3556787	68	#N/A	3556855
Rv3196A	+	3566981	85	3567029	3566896
Rv3646c	-	4087424	167	4087380	4087257
Rv3659c	-	4098011	17	#N/A	4097994
Rv3818	+	4282411	38	4282412	4282449
Rv3846	+	4320065	639	4320619	4320704

SNP poisitions highligeted in red were not considered for further analysis because they were not withing 500bp of the start of a gene

* Sub-lineage 4 intergenic (non-coding) SNPs (SL-4_IG_SNP)

The relative distances of the ncSNPs of sub-lineage 4 downstream from a gene start site is shown in Figure 3.24.

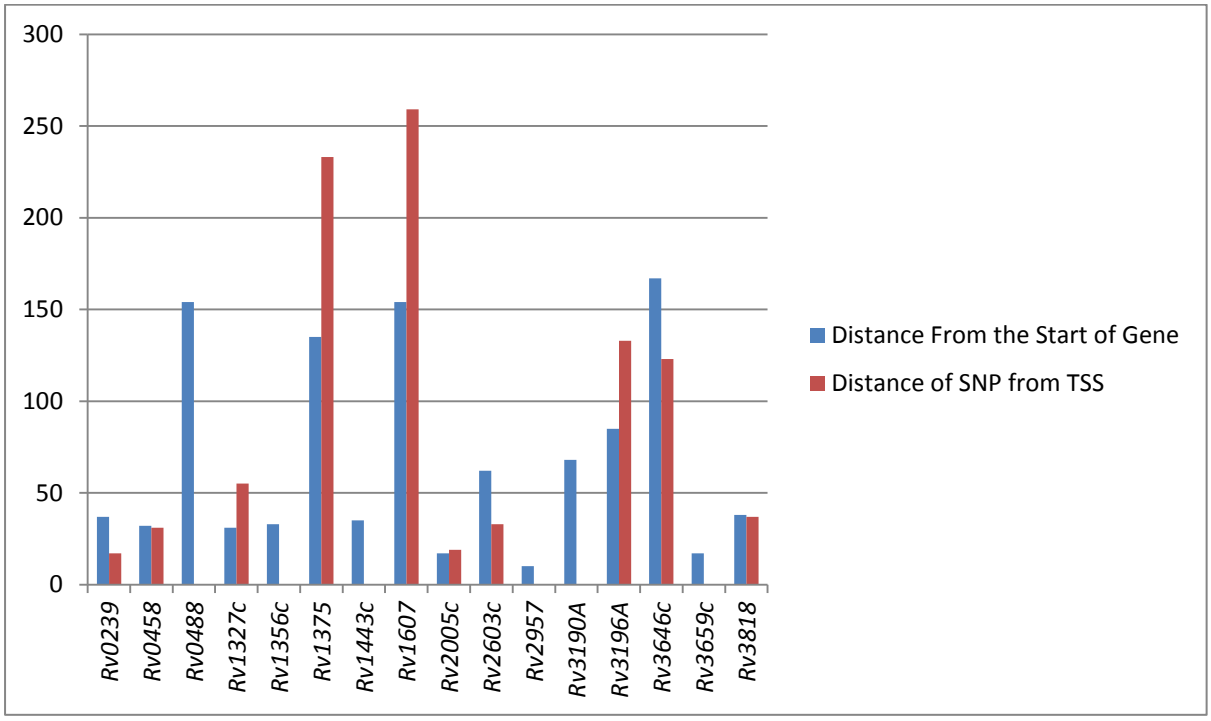


Figure 3.24: Graph illustrating the distance of ncSNP from the start of adjacent gene and from the genes reported transcription start site (TSS) for sub-lineage 4.

The study of sub-lineage 5 revealed 14 unique ncSNPs of which 2 SNPs were more than 500bp from the start of the nearest gene. Furthermore, 6 TSS were identified among 12 genes having ncSNPs before the gene start site as shown in Table 4.10.

Table 3.10: Sub-lineage 1distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS)

Nearest Gene	Orientation	SL-5_IG_SNP	Distance from Gene	TSS	Gene Start Position
Rv0071	+	79463	23	#N/A	79486
Rv0302	+	364514	91	364523	364605
Rv0453	-	543170	4	#N/A	543174
Rv0473	+	563522	42	563567	563564
Rv0522	+	612904	134	612908	613038
Rv0547c	-	640030	1114	#N/A	638916
Rv1215c	-	1359456	12	1359451	1357759
Rv1774	+	2007813	19	#N/A	2007832
Rv1855c	-	2104140	33	2104107	2104107
Rv1986	-	2230944	-933	2229988	2230011
Rv2309A	-	2583011	34	#N/A	2583045

Rv2878c	-	3189580	183	#N/A	3189397
Rv3129	+	3494493	167	#N/A	3494660
Rv3428c	-	3846269	299	#N/A	3845970

*SNP positions highlighted in red were not considered for further analysis because they were not within 500bp of the start of a gene

The relative distances of the ncSNPs downstream from gene start site and its associated TSS are illustrated in figure 3.25.

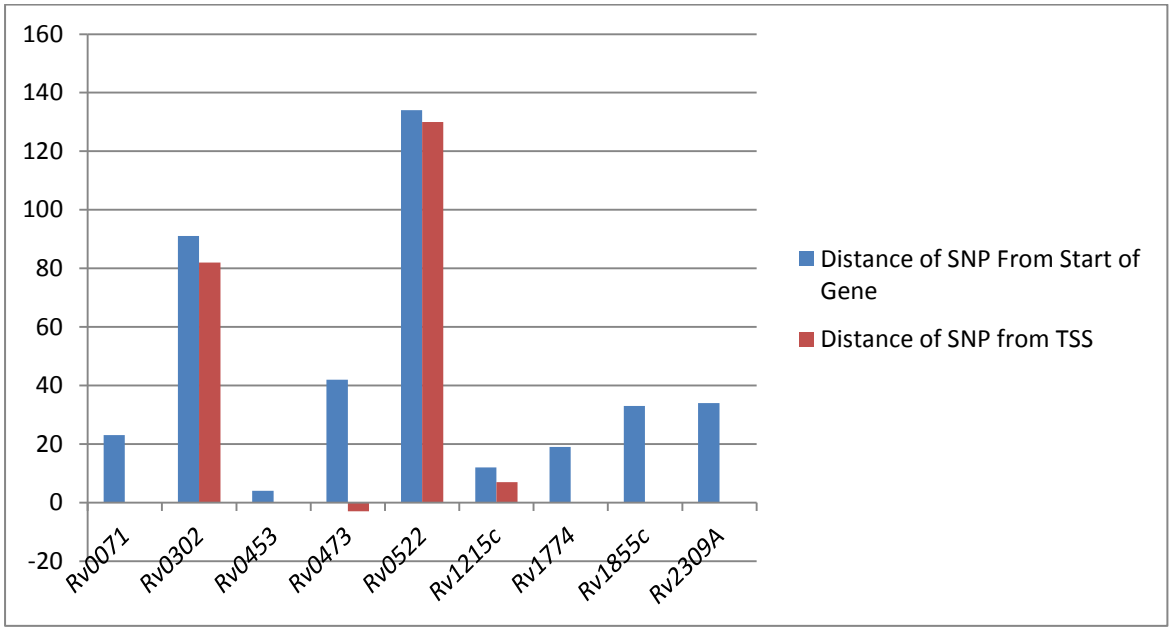


Figure 3.25: Graph illustrating the distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

Sub-lineage 6 had a total of 15 unique ncSNPs of which 12 were within 500bp downstream of a gene start site in this study.

Table 3.11: Sub-lineage 1distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

Nearest Gene	Orientation	SL-6_IG_SNP	Distance from Gene	TSS	Gene Start Position
Rv0147	+	173184	54	173231	173238
Rv0885	+	982687	64	982724	982762
Rv1051c	-	1174998	298	#N/A	1174700
Rv1184c	-	1325619	1	1325688	1325618
Rv1261c	-	1409474	-459	1409933	1409933
Rv1566c	-	1774641	4	1774722	1774637
Rv1753c	-	1984778	3	1984870	1984775

Rv1918c	-	2170860	248	2170683	2170612
Rv1984c	-	2228937	376	2228561	2228561
Rv1994c	-	2238000	16	2237984	2237984
Rv2560	+	2878550	1525	2880020	2880075
Rv3033	+	3393258	122	3393350	3393380
Rv3191c	-	3558616	271	#N/A	3558345
Rv3813c	-	4278378	-837	#N/A	4278394
Rv3826	+	4299610	202	4299810	4299812

*SNP positions highlighted in red were not considered for further analysis because they were not within 500bp of the start of a gene

A total of 13 ncSNP were within 500bp downstream of the start site of a gene of which 11 had a previously described TSS. A total of 3 ncSNPs were situated upstream of a TSS (*Rv1918c*, *Rv1984c* and *Rv1994c*) and are illustrated in Figure 3.26

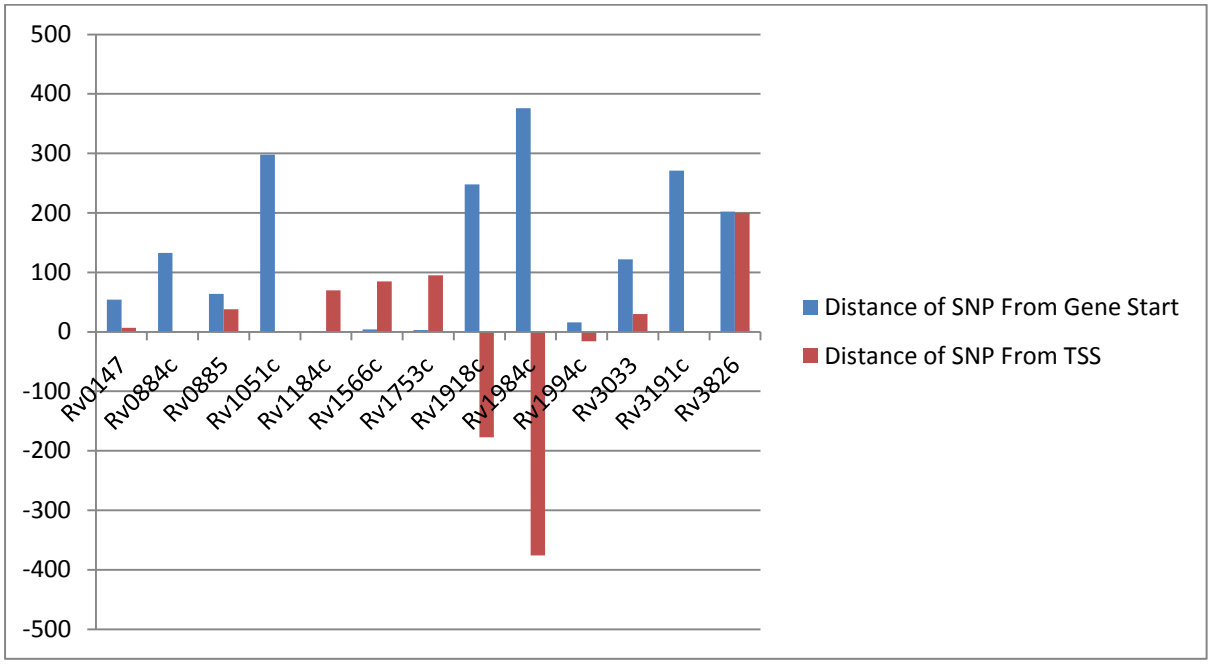


Figure 3.26: Graph illustrating the distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

Sub-lineage 7, which was reported to be responsible for most of the tuberculosis cases in Cape Town, had 9 unique ncSNPs of which 2 were situated more than 500bp from the start of their nearest gene as shown in Table 4.12. Six genes which had ncSNPs downstream of the start site of the nearest gene had TSSs established by Cortes *et al.* (2013).

Table 3.12: Sub-lineage 7: distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

Nearest Gene	Orientation	Sub-Lineage 7 Non-coding SNPs	Distance from Gene	Transcription Start Site	Gene Start Position
Rv0104	+	122205	112	#N/A	122317
Rv0687	+	786965	134	#N/A	787099
Rv1173	+	1302818	113	1302855	1302931
Rv2344c	-	2623778	26	#N/A	2623752
Rv3124	+	3490438	-932	#N/A	3489506
Rv3453	+	3874206	198	3492027	3874404
Rv3742c	-	4193344	99	3874240	4193245
Rv3770A	+	4215793	-270	4193314	4215881
Rv3898c	-	4384127	142	#N/A	4383985

SNP positions highlighted in red were not considered for further analysis because they were not within 500bp of the start of a gene

All the ncSNPs within 500bp downstream of gene start site were also situated downstream of TSS where these were previously described as illustrated in Figure 3.27

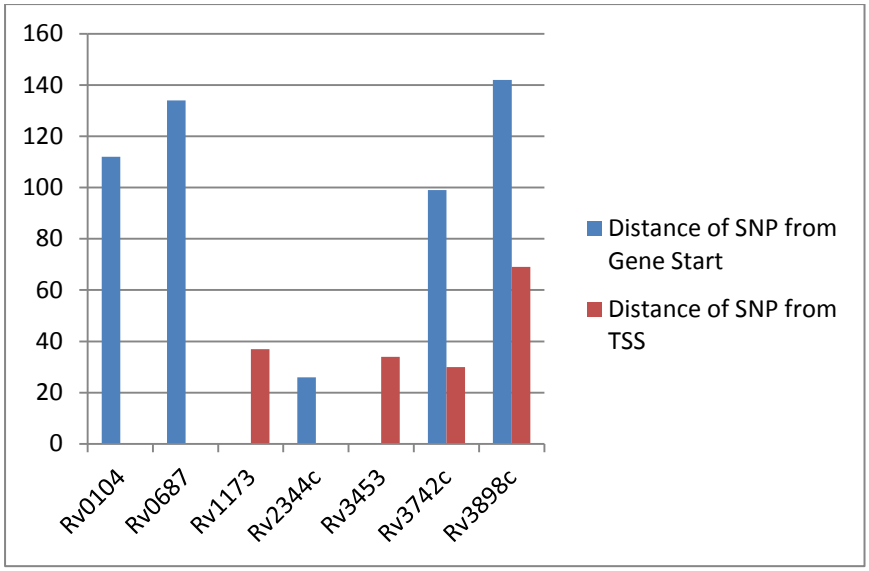


Figure 3.27:Graph illustrating the distance of ncSNP from the start of adjacent gene and reported transcription start site (TSS).

3.16 Comparison between hypo- and hypervirulent M. tuberculosis Beijing sub-lineage 7

Whole genome sequencing was done in order to identify the genomic mutations which could account for the observed hypo-virulent and hyper-virulent phenotypes, observed in two closely related Beijing sub-lineage 7 strains (SAWC 5527 and SAWC 507). Analysis of these genome sequences showed that they share 1694 SNPs with respect to *M. tuberculosis* H37Rv when considering SNPs identified by all 3 mapping algorithms. The hyper-virulent SAWC 5527 had 56 unique SNPs when compared to the hypo-virulent SAWC 507 strain which in turn had 28 unique SNPs encompassing sSNPs, nsSNPs and ncSNPs. The number of unique SNP decreased to 15 for both strains after excluding SNPs in repetitive regions such as *pe/ppe* genes and insertion sequences in the coding regions (Table 3.13). Additionally, there were also 7 unique intergenic SNPs in the hyper-virulent strain and 1 in the hypo-virulent strain.

Table 3.13: Number of unique and total SNPs from overlap of 3 genome mappers for sub-lineage 7 hypo-virulent and hypo-virulent strains grouped by Tuberculist functional category.

		Unique SNPs Grouped by Functional Category							
Strains	**Total SNPs**	**Cell wall and cell processes**	**Information Pathways**	**Conserved hypotheticals**	**Intermediary metabolism and respiration**	**Lipid metabolism**	**Virulence, detoxification, adaptation**	**Regulatory proteins**	**Intergenic SNPs**
SAWC507 (hypo-virulent)	1721	3	0	1	7	2	0	2	1
SAWC5527 (hyper-virulent)	1750	5	1	3	4	1	1	0	7

3.17 Functional effect of nsSNPs

A more detailed analysis looking at the functional effect of SNPs was done using 4G Sorting Intolerant from Tolerant (SIFT) (http://sift.bii.a-star.edu.sg/) algorithm which is the updated version of SIFT. The 4G SIFT predicted functional nsSNPs of the hypo-virulent and hyper-virulent strains were subsequently compared to their previously described whole proteome results. The proteomic investigations had revealed that there was an over representation of regulatory and cell wall proteins when considering overly abundant proteins between the hypo-virulent and hyper-virulent strains (de Souza *et al*., 2010). The number of genes containing 4G SIFT deleterious nsSNPs in each functional category was nearly identical between the hypo-virulent and hyper-virulent strains, as shown in Table 3.14. Additionally, there was agreement in the total number of unique SNPs between those identified from 3 genome mappers in Table 3.13 and those identified by SIFT in Table 3.14. Among the exceptions was 1 functional nsSNP in Rv0988, which was found only in the hyper-virulent strain amongst 25 nsSNPs associated with cell wall and cell processes. The hypo-virulent strain had a total of 2 predicted deleterious SNPs and the hyper-virulent a total of 6 as indicated in Table 3.14.

Table 3.14: Number of total SNPs (tolerated and deleterious) and deleterious nsSNPs predicted by 4G SIFT for sub-lineage 7 hypo-virulent and hypo-virulent strains grouped by Tuberculist functional category.

Tuberculist Functional Category	**SAWC 507 (Hypo-virulent)**	**SAWC 5527 (Hyper-virulent)**	**Unique SAWC 507 (Hypo-virulent)**		**Unique SAWC 5527 (Hyper-virulent)**	
			Total Unique SNPs	Deleterious SNPs	Total Unique SNPs	Deleterious SNPs
Cell Wall and Cell Processes	24	25	3	0	5	1
Conserved Hypothetical proteins	25	26	1	0	3	1
Information Pathways	16	17	0	0	1	
Intermediate Metabolism and Respiration	33	34	7	0	4	2
Lipid Metabolism	14	14	2	2	2	1
Regulatory proteins	8	8	2	0	0	0
Virulence, Detoxification, Adaptation	3	3	0	0	1	1

The results of the 4G SIFT predicted functional SNPs were compared to the proteomics expression data of the hypo-virulent and hyper-virulent virulent strains to establish whether there was any correlation between these data sets. The proteins which were uniquely present in the hypo-virulent and hyper-virulent strains were specifically looked at in relation to the unique 4G SIFT predicted high confidence functional SNPs as depicted in Table 3.15.

The nsSNP in Rv0988 has previously been reported to be part of the Rv0986-8 operon. The operon has been hypothesized to be involved in *M. tuberculosis* virulence and host-pathogen interactions (Liu *et al*., 2014; Rosas-Magallanes *et al*., 2006). Additionally, Rv0214 which is associated with lipid metabolism had a deleterious nsSNP in the hyper-virulent strain but not in the hypo-virulent strain and was supported by proteomics. A second lipid metabolism functional mutation was also identified in Rv2934 only in the hypo-virulent strain. This was expressed in the hyper-virulent strain from proteomics investigations but not in the hypo-virulent strain. Rv2934 has been identified in *M. tuberculosis* guinea pig infection studies at day 30 and 90 in a study by Kruh *et al*., 2010. The Rv3542c (hypothetical protein functional category) predicted functional SNP in the hyper-virulent strain had no

corresponding proteomics expression. In contrast, the hypo-virulent strain didn't have a predicted functional SNP in Rv3542 and had associated protein expression. Studies by others have however reported that Rv3542 is required for survival in primary murine macrophages (Rengarajan *et al*., 2005).

Table 3.15: Comparison of Unique 4G SIFT High Confidence Predicted Functional SNPs to Unique Protein Expression Data of Hypo-Hyper Virulent Strains

Unique 4G SIFT Predicted to Have Functional Effect	**Presence/absence of 4G SIFT predicted SNP to have functional effect SAWC 507 (Hypo-virulent)**	**Presence/absence of 4G SIFT predicted SNP to have functional effect SAWC 5557 (Hyper-virulent)**	**SAWC 507 (Hypo-virulent) proteomics presence/absence**	**SAWC 5527 (Hyper-virulent) Proteomics presence/absence**
Rv0169	Absent	Present	Absent	Absent
Rv0214	Absent	Present	Present	Absent
Rv0753c	Absent	Present	Absent	Absent
Rv0988	Absent	Present	Absent	Absent
Rv1318c	Absent	Present	Absent	Absent
Rv3542c	Absent	Present	Present	Absent
Rv1760	Present	Absent	Absent	Absent
Rv2934	Present	Absent	Absent	Present

An analysis of deleterious 4G SIFT predicted nsSNPs with low confidence revealed that there was no such mutation in SAWC 507. There was however a total of 12 tolerated functional nsSNPs identified in the hypo-virulent SAWC 507 of which two SNPs had a regulatory functional grouping annotation. These were *Rv3197A* (whiB7) and *Rv3220c*. The hyper-virulent SAWC 5527 on the other hand had 1 deleterious 4G SIFT predicted nsSNPs with low confidence in *Rv0213c*. There was however no evidence of proteomics expression associated with this mutation.

3.18 Large Duplication events in the hyper-hypo virulent strains analysis

Large genomic duplication analysis was undertaken for the hypo-hyper virulent

strains using BEDTOOLS. Table 3.16 and 3.17 shows genomic dupications (>500bp) in the hypo-virulent and hyper-virulent strains with high sequence coverage, excluding repeat regions.

Table 3.16: Regions of high sequence coverage in the hypo-virulent strain. The shaded area in blue and purple represent the continuous region of high sequence coverage

Reference	**Start**	**End**	**Size**	**High Coverage Total Size**
gi|444893469|emb|AL123456.3|	25919707	2591716	269	765
gi|444893469|emb|AL123456.3|	3793294	3791224	3	**2074**

Table 3.17: Regions of high sequence coverage in the hyper-virulent strain. The shaded area in blue and purple represent the continuous region of high sequence coverage

Reference	Start	End	Size	**High Coverage Total Size**
gi|444893469|emb|AL123456.3|	2592210	2591708	232	**735**
gi|444893469|emb|AL123456.3|	3793267	3791471	24	**1821**

3.19 Indel analysis of the hypo-hyper virulent strain

Three small indels were identified in the hypo-strain whilst the hyper–virulent strain had 20 indels as shown in Tables 3.18 and 3.19. The indel analysis of the hyper-hypo virulent in comparison to the proteomics data revealed that there was no direct effect of the small indels detected in the hypo and hyper virulent strains with respect to protein expression in-vitro.

Table 3.18: Small indels identified in the hypo-virulent strain

POS	REF	MUTATION BWA	TYPE	NAME	PRODUCT	FUNCTIONAL CATAGORY	FUNCTION
1273250	G	GA	CDS MUTATION	Rv1145	Probable conserved transmembrane transport protein MmpL13a	cell wall and cell processes	Unknown. Thought to be involved in fatty acid transport.
3610391	A	AC	CDS MUTATION	Rv3234c	Putative triacylglycerol synthase (diacylglycerol acyltransferase) Tgs3	lipid metabolism	May be involved in synthesis of triacylglycerol
4170964	G	GA	CDS MUTATION	Rv3725	Possible oxidoreductase	intermediary metabolism and respiration	Function unknown; probably involved in cellular metabolism.

Table 3.19: Small indels identified in the hyper-virulent strain

POS	REF	MUTATION BWA	TYPE	NAME	PRODUCT	FUNCTIONAL CATAGORY	FUNCTION
162151	GT	G	CDS MUTATION	Rv0134	Possible epoxide hydrolase EphF (epoxide hydratase) (arene-oxide hydratase)	virulence, detoxification, adaptation	Thought to be involved in detoxification reactions following oxidative damage to lipids [catalytic activity: an epoxide + H(2)O = a glycol].
194305	C	CGG	CDS MUTATION	Rv0165c	Probable transcriptional regulatory protein Mce1R (probably GntR-family)	regulatory proteins	Involved in transcriptional mechanism
234496	C	CGT	CDS MUTATION	Rv0197	Possible oxidoreductase	intermediary metabolism and respiration	Function unknown; probably involved in cellular metabolism.
670386	AC	A	CDS MUTATION	Rv0576	Probable transcriptional regulatory protein (possibly ArsR-family)	regulatory proteins	Involved in transcriptional mechanism.
1370851	GT	G	INTERGENIC				
1387656	CG	C	CDS MUTATION	Rv1244	Probable lipoprotein LpqZ	cell wall and cell processes	Unknown

1625332	C	CGGT	CDS MUTATION	Rv1446c	Putative OXPP cycle protein OpcA	intermediary metabolism and respiration	May be involved in the functional assembly of glucose 6-phosphate dehydrogenase
2161343	G	GT	CDS MUTATION	Rv1915	Probable isocitrate lyase AceAa [first part] (isocitrase) (isocitratase) (Icl)	intermediary metabolism and respiration	Involved in glyoxylate bypass, an alternative to the tricarboxylic acid cycle [catalytic activity: isocitrate = succinate + glyoxylate].
2403603	TG	T	CDS MUTATION	Rv2143	Conserved hypothetical protein	conserved hypotheticals	Unknown
2406842	AC	A	CDS MUTATION	Rv2147c	Conserved hypothetical protein	conserved hypotheticals	Unknown
2564368	G	GC	CDS MUTATION	Rv2293c	Conserved hypothetical protein	conserved hypotheticals	Unknown
3441938	GC	G	CDS MUTATION	Rv3079c	Conserved protein	conserved hypotheticals	Function unknown
3590686	G	GC	INTERGENIC				
3610391	A	AC	CDS MUTATION	Rv3234c	Putative triacylglycerol synthase (diacylglycerol acyltransferase) Tgs3	lipid metabolism	May be involved in synthesis of triacylglycerol
4130610	GC	G	CDS MUTATION	Rv3689	Probable conserved transmembrane protein	cell wall and cell processes	Unknown
4170964	G	GA	CDS MUTATION	Rv3725	Possible oxidoreductase	intermediary metabolism and respiration	Function unknown; probably involved in cellular metabolism.
4264381	G	GC	INTERGENIC				
4327585	AG	A	CDS MUTATION	Rv3855	Transcriptional regulatory repressor protein (TetR-family) EthR	regulatory proteins	Regulates negatively the production of ETHA. Induced ETH resistance when overexpressed in Mycobacterium tuberculosis.
4338595	GC	G	INTERGENIC				

3.20 Region of difference deletion analysis

In this study we identified regions of difference (RD)s by analysing regions which exhibited no alignments of reads in whole genome mapping using H37Rv as a reference. The Beijing RD lineage defining marker RD105 depicted in Figure 3.28 was identified in all of the Beijing strains in this study. Results of the other Beijing RDs as well as RD149 and RD152 which are also found in other lineages is shown in Table 3.20.

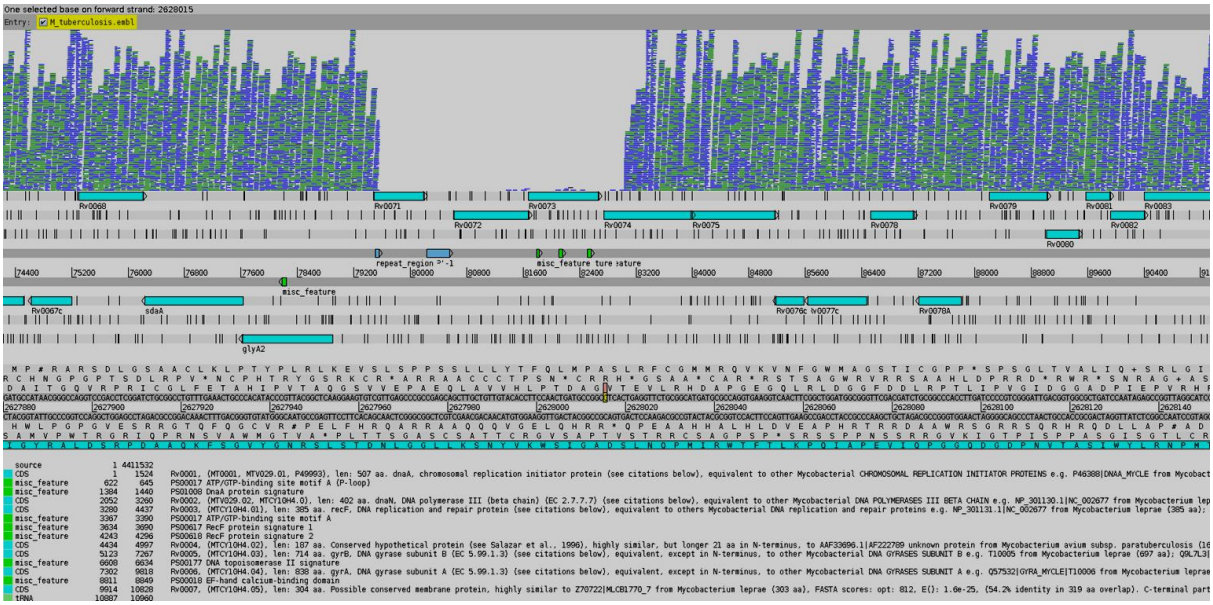


Figure 3.28: RD 105 Deletion which is a *Mycobacterium tuberculosis* Beijing lineage defining marker.

The presence of RD149 and RD152 deletions have been implicated in the virulence of strains in studies by other (Kanji *et al*., 2011b).

Table 3.20: RD analysis showing known deletion markers (green highlights indicate region deleted while red highlights region present).

Beijing sub-lineage	Regions of Difference												
	RD105				RD150		RD181		RD142		RD152	RD149	RD207
	79567 - 83034				1896862 -1899349		2535429 -2536140		1332182 -1335033		1986636 - 199862	1,619,027 - 1,633,610	1779264 - 178851
	Genes												
	Rv0071	Rv0072	Rv0073	Rv0074	Rv1672c	Rv1673c	Rv2262c	Rv2263	sigI	Rv1190-Rv1192	plcD-Rv1765c	PGRS26 - PE PGRS27	Rv1573-Rv1588c
	Function												
	Unknown	Thought to be involved in active transport of glutamine across the membrane	Thought to be involved in active transport of glutamine across the membrane	Unknown	Thought to be involved in active transport across the membrane	Unknown	Function unknown; Thought to be involved in lipid metabolism	Oxidative-reduction	Oxidises proline to glutamate for use as a carbon and nitrogen source	Unknown	-Hydrolyzes shingomyelin to phosphatidylcholine -Required for transposition of IS*6110* - Hydrolysis of cutin -Unkwon - May be involved in synthesis of triacylglycerol	- Unknown - Enzyme may serve as scavenger - Pentose phosphate pathway - May be involved in the functional assembly of glucose 6-phosphate dehydrogenase - Translodase	- Unknown - Integration of phiRv1 into chromosome .
Beijing Sub-lineage 1	X				-		-		-		X	-	X
Beijing Sub-lineage 2	X				-		X		-		X	-	X
Beijing Sub-lineage 3	X				-		X		-		X	-	X
Beijing Sub-lineage 4	X				-		X		-		X	X	X
Beijing Sub-lineage 5	X				-		X		-		X	-	X
Beijing Sub-lineage 6	X				-		X		-		X	-	X
Beijing Sub-lineage 7	X				X		X		-		X	-	X

We further looked for RDs that were unique in a sub-lineage or present in some but not all of the strains analysed in this study. The unique in-silico identified RD for Beijing sub-lineage are illustrated in Figure 3.29 whilst those common to sub-lineages 4,5,6 and 7 are shown in Figure 4.30. Table 3.21 shows the in-silico RDs identified in this study.

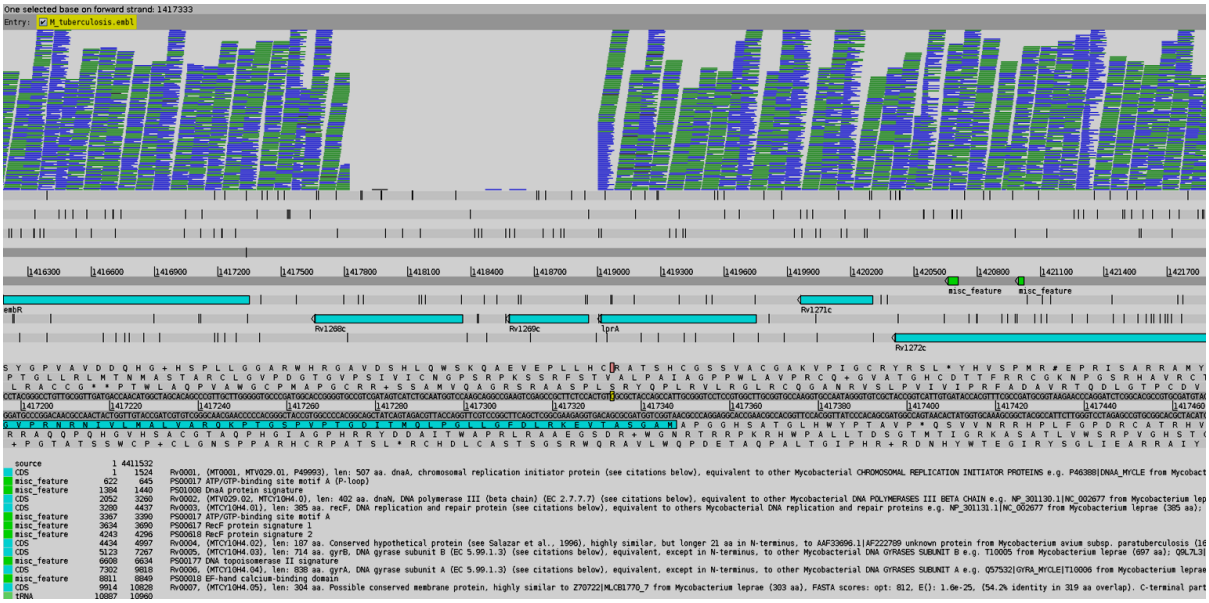


Figure 3.29: Sub-lineage 4 Mycobacterium tuberculosis Beijing lineage defining RD marker.

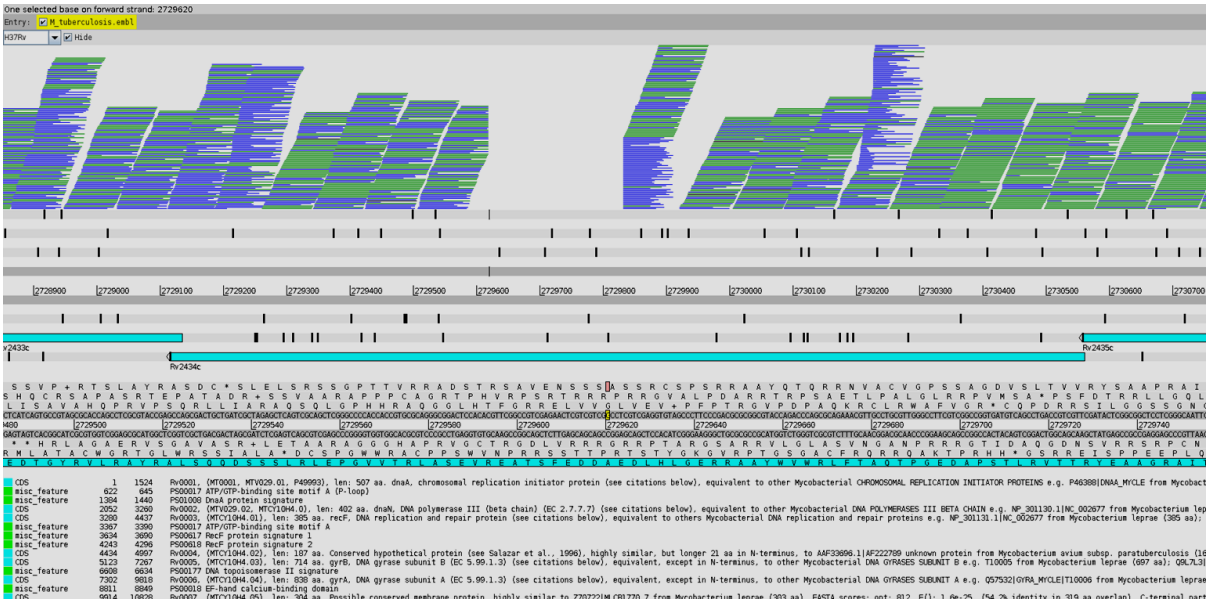


Figure 3.30 Sub-lineage 4,5,6 and 7 Mycobacterium tuberculosis Beijing lineage defining RD marker.

Table 3.21: Region of difference analysis showing identified potential deletion markers in this study with green highlights showing presence of deletion and red indicating absence of deletion.

	Unique Regions of Difference												
Beijing Sub-lineage	**RD1_1**	**RD3_1**				**RD4_1**		**RD5_1**	**RD6_1**			**RD7_1**	**RD_4-5-6-7**
	2,128,379-2,129581	2,116,343- 2,121.097				1,417,827- 1,419,002		3,730,768-3,732,115	2,628,512- 2,634,021			3,378,553-3,379,024	2,729,620-2,729,832
	Rv1878	Rv1867	Rv1868	Rv1869c	Rv1870c	Rv1268c	Rv1269c	Rv3343c	Rv2350c	Rv2351c	Rv2352c	Rv3019c	Rv2434c
	Involved in glutamine biosynthesis	Function unknown, but supposed involvement in lipid degradation	Function unknown	Function unknown; probably involved in cellular metabolism	Function unknown	Function unknown	Function unknown	Function unknown	Hydrolyzes sphingomyelin in addition to phosphatidyl choline	Hydrolyzes sphingomyelin in addition to phosphatidylch oline	Function unknown	Function unknown	Function unknown
Beijing Sub-lineage 1	X	-				-		-	-			-	
Beijing Sub-lineage 2		X											-
Beijing Sub-lineage 3	-	X				-		-	-			-	-
Beijing Sub-lineage 4	-	-				X		-	-			-	X
Beijing Sub-lineage 5	-	-				-		X	-			-	X
Beijing Sub-lineage 6	-	-				-		-	X			-	X
Beijing Sub-lineage 7	-	-				-		-	-			X	X

3.21 Hybrid assembly

Whole genome de novo assembly was done using a hybrid assembly method for the 2 sub-lineage 7 Beijing strains (SAWC 507 and SAWC 5527). The hybrid de novo assembly utilizes the advantages of long sequence reads of PacBio sequencing and the paired-end information of Illumina paired-end (PE) sequence reads. Hybrid assemblies using both PacBio and Illumina sequences were done using CELERA and MIRA genome assemblers. Statistics on the assembled genome are given in Table 3.22 and these show that despite having the same input files, the results generated by CELERA and MIRA differ in terms of number of contigs constructed, total bases incorporated and the size of the largest contig in the assembly.

Table 3.22: Statistics on CELERA and MIRA genome assemblies

Strain	**Total Scaffolds**		**Total contigs in scaffolds**		**Total bases in scaffolds**		**Max contig length**		**N50 contig bases****	
	CELERA	*MIRA	CELERA	***MIRA (Not in Scaffolds)	CELERA	MIRA	CELERA	MIRA	CELERA	MIRA
5527	105	N/A	108	186	4511081	4598636	406235	377348	160315	148300
507	67	N/A	77	123	4434229	4482087	371796	541525	170676	102492

*Large contigs greater than or equal to 500bp.

** A N50 contig is sought in order to estimate the accuracy of an assembly. The N50 contig is the contig of least length of a set of minimum number of contigs whose combined length accounts for 50% of a genome assembly (Miller J.R *et al et al* 2010).

*** MIRA does not output scaffolds but only contigs

3.22 De novo assembly gap closure

Illumina paired-end read information was used to close gaps between a set of contigs. For example, if a repeat region falls in between a read pair separated by a known insert size distance, contigs are joined together and gaps subsequently closed as is illustrated in Figure 3.31. The number of gaps closed among the contigs using paired-end Illumina reads are given in Table 3.23. The number of gaps that needed closing as determined by ABACAS was 47 and 91 for CELERA and MIRA assemblies for the SAWC507 hypo-virulent strain respectively. For the SAWC5527 hypo-virulent strain, these were 56 and 129 for CELERA and MIRA assemblies respectively.

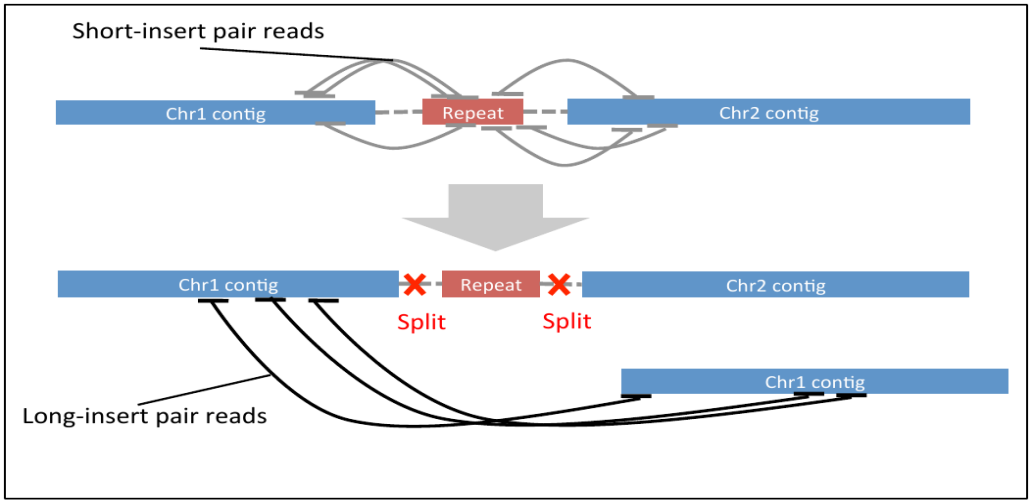


Figure 3.31: Paired reads resolving a gap between to contigs with repeat sequences within them as a result of at least one pair of the reads not overlapping the repeat sequence.

Table 3.23: Number of gaps closed with PLATATANUS in CELERA and MIRA assemblies of hypo-virulent and hyper-virulent Beijing sub-lineage 7 strains.

Strain	Total of Gaps Closed	
	CELERA	MIRA
SAWC507 (Hypo-virulent)	9	1
SAWC5527 (Hyper-virulent)	4	0

3.23 ABACAS contig ordering of genome assemblies

The de novo assemblies of the hypo-virulent and hyper-virulent strains were ordered against the reference H37Rv to enable better comparative analysis and viewing of the assemblies using MAUVE. This however can result in a number of contigs being removed from the initial de novo assemblies and these are indicated in Table 3.24.

Table 3.24: ABACAS statistics showing the number of unused contigs

Strain	No. of unused contigs	
	CELERA	MIRA
SAWC507 (Hypo-virulent)	0	5
SAWC5527(Hyper-virulent)	3	6

3.24 MAUVE view of assembly

MAUVE aligner was used to align the assembled genomes to *M. tuberculosis* H37Rv reference genome and used to visualize the assembly. Using this approach the presence of the Beijing lineage RD105 large deletion illustrated in Figure 3.32 was verified in both the hyper and hypo-virulent Beijing lineage strains.

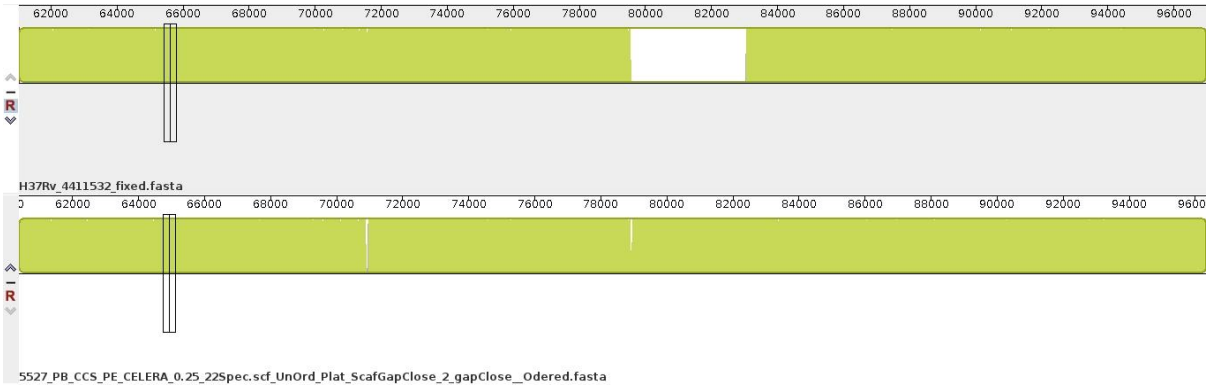


Figure 3.32: Section of genome showing Beijing RD105 large deletion in the hyper-virulent SAWC 5527 de novo assembly genome constructed using CELERA.The top genome in illustration is the reference H37Rv and the bottom genome is SAWC 5525. The white region in the reference highlights area which is different or absent in the genome of SAWC 5527.

The full assemblies of the hypo-virulent and hyper-virulent strains using MIRA and CELERA genome assemblers and aligned to H37Rv using MAUVE are illustrated in Figures 3.33, 3.34, 3.35 and 3.36. The fewer the number of coloured blocks represent a better aligning of assemblies by MAUVE. The coloured blocks appearing below the horizontal axis in both SAWC 507 and SAWC 5527 strain when aligned to the reference H37Rv depicts an inversion event with respect to the reference H37Rv. There was lack of agreement in the number and locations of inversion event arising when using MIRA and CELERA to construct the de novo assemblies on the same data sets.

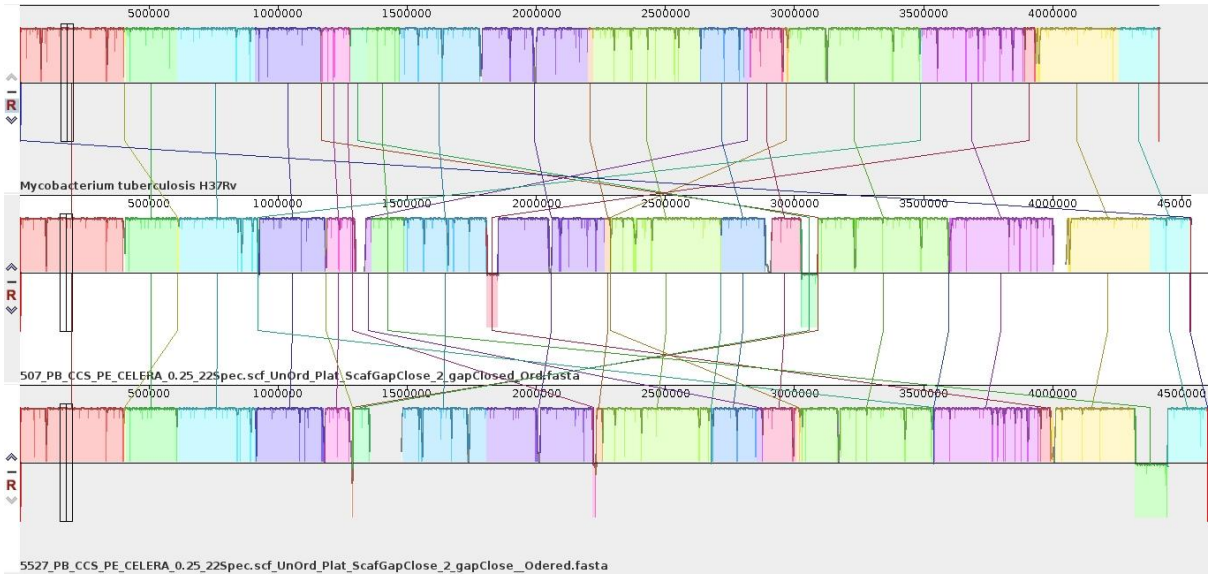


Figure 3.33: SAWC 507 de novo assembly by CELERA with coloured blocks depicting contiguous stretches of assemblies aligned to H37Rv. The top genome in illustration is the reference H37Rv and the bottom genome is SAWC 507. The lines connecting blocks highlight which blocks between the the two genomes are similar whilst the linear numbers represent base positions along the genome of SAWC 507. The coloured blocks appearing below the horizontal axis in the SAWC 507 strain depicts an inversion event with respect to the reference H37Rv.

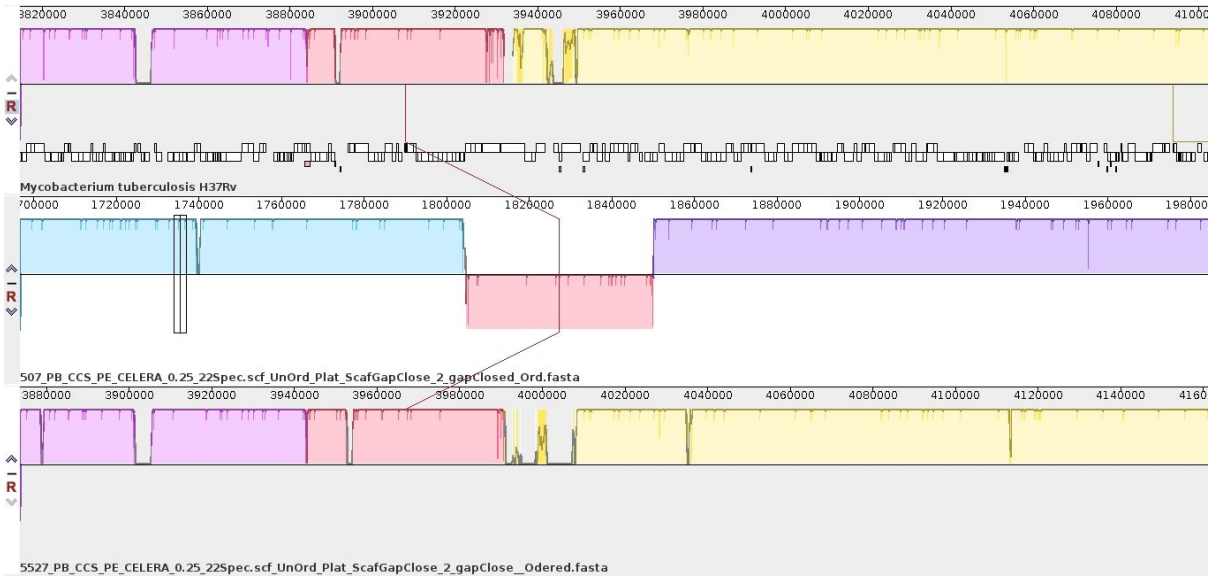


Figure 3.34: SAWC 5527 de novo assembly by CELERA with coloured blocks depicting contiguous stretches of assemblies aligned to H37Rv. The top genome in illustration is the reference H37Rv and the bottom genome is SAWC 5527. The purple coloured block appearing below the horizontal axis in the SAWC 5527 strain depicts an inversion event with respect to the reference H37Rv.

Figure 3.35: SAWC 507 and SAWC 5527 *de novo* assembly by CELERA. Coloured blocks depicting contiguous stretches of assemblies aligned to H37Rv. The top genome in illustration is the reference H37Rv, the middle genome SAWC 507 and the bottom genome is SAWC 5527. The coloured block appearing below the horizontal-axis in the SAWC 5527strain depicts an inversion event with respect to the reference H37Rv.

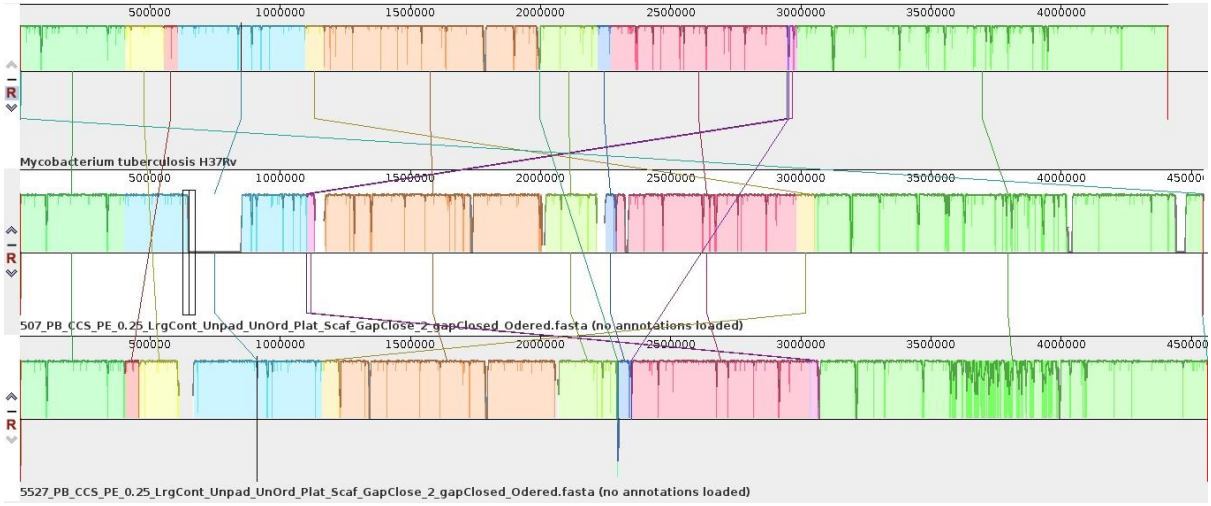


Figure 3.36: SAWC 507 and SAWC 5527 de novo assembly by MIRA with coloured blocks depicting contiguous stretches of assemblies aligned to H37Rv. The top genome in illustration is the reference H37Rv, the middle genome SAWC 507 and the bottom genome is SAWC 5527.

MAUVE produced fewer coloured blocks (contiguous stretches of assemblies aligned to H37Rv) with MIRA assemblies than it did for CELERA for both the hypo-virulent and hyper-virulent strains in this study. Annotation of the assembled genome was subsequently done using RAST.

3.25 Genome Annotation using RAST

Annotation of the assembled genomes was carried out using the RAST annotation tool (http://rast.nmpdr.org) (Aziz *et al* 2008). Both the ordered and unordered contigs of the assembled hypo-virulent and hyper-virulent sub-lineage 7 genomes were uploaded for annotation. The number of features (protein coding sequence) identified for each assembled genome is shown in Table 3.25.

Table 3.25: Number of contigs and associated RAST annotated features for ordered and non-ordered MIRA and CELERA assemblies.

		Number of Contigs		**Number of Features***	
Strain	**Genome Assembler**	**Ordered Contigs**	**Unordered Contigs**	**Ordered Contigs**	**Unordered Contigs**
SAWC 507	MIRA	91	127	4182	4462
SAWC 507	CELERA	52	77	4292	4387
SAWC 5527	MIRA	130	191	4402	4584
SAWC 5527	CELERA	58	108	4372	4472

* Protein coding sequence

The number of features annotated by RAST was more when using contigs not ordered with ABACAS for both MIRA and CELERA assemblies.

3.26 IS6110 location in assembled genomes

Following de novo assembly of the 2 strains and ordering of contigs using ABACAS, locations of the transposon element IS*6110* were determined via sequence matching of the beginning and end of IS*6110* as indicated in the materials and methods. Results of the sequence search matches and subsequent NCBI BLAST searches of sequences adjacent to identified insertion points are shown in Table 3.26. In addition to searching for insertion points in the ordered contigs using ABACAS, searches were also done in the unordered contigs as there were instances where ABACAS did not use some contigs sections in its ordered genome output as depicted in Table 3.20.

This in silico analysis showed that the hypo-virulent strain (SAWC 5527) had 25 IS*6110* insertion points where as the hyper-virulent strain (SAWC 5527) had 24 IS*6110* insertion points. Two insertion points for IS*6110* were unique to the hypo-virulent strain at positions 1892 and 3711737 relative to H37Rv. The insertion at position 3711737 in the hypo-virulent strain had an associated inversion event previously reported by (Shitikov *et al*., 2014) in the CELERA assembly. The hyper-virulent strain had 1 unique insertion point at position 3076124 with respect to *M. bovis* AF2122/97 complete genome.

There was agreement between the IS*6110* search results of MIRA and CELERA contigs with the exception of positions 3076124 (with respect to *M. bovis* AF2122/97 complete genome shown in Table 3.26) and 3711737. The in-silico insertion points identified in the 2 Beijing strains have previously been reported by Alonso *et al*., (2013) Beijing lineage studies with the exception of insertions at position 1892 in H37Rv and position 3076124 with respect to *M. bovis* AF2122/97 complete genome.

Table 3.26: Location of IS*6110* in the ordered and non-ordered MIRA and CELERA assemblies identified by searching for the beginning and end of IS*6110* in the genome assemblies. The begining of the IS*6110* sequence in the forward orientation is denoted FWD_BEGIN in the table whilst the end is denoted FWD_End. For the reverese orientation of IS*6110*, REV_BEG denotes the begining of the IS*6110* sequence in the reverse orientation in the table whilst the end is denoted REV_End. Columns indicated with an * contain the location of insertion of IS*6110*, part of IS*6110* used in search and the contig number in assembly containing the aforementioned IS*6110* match.

No.	Name of Reference	Present in MIRA Assembly	Present in CELERA Assembly	Gene	Title	Present in SAWC507 Ordered	*Present in MIRA 5SAWC07 Unodered	*Present in CELERA SAWC507 Unodered	Point of Insertion	Present in SAWC5 527 Ordered	*Present in MIRA SAWC 5527 Unordered	*Present in CELERA SAWC5527 Unodered	Point of Insertion
FWD_BEGIN: TGAACCGCCCCGGTGAGTCCGGAGACTCTCTGATCTGAGACCTCAGCCGGCGGCTGG													
1	H37Rv-emb|AL123456.3|	YES	YES	Rv0756c	Unknown	YES	851479 Rev-End Scaf 7	851476 Fwd-End Scaf 22 851476 Rev-Beg Scaf 61 851479 Rev-End Scaf 28	851478 (CELERA Only)	YES	851478 Fwd-Beg Scaf 19 GTA 851476 Rev-Beg Scaf 9 TAC	851478 Fwd-Beg Scaf 95 851476 Rev-Beg Scaf 85 ATAC	851478 (CELERA Only)
2	H37Rv-emb|AL123456.3|	YES	YES	Rv0962c	Possible lipoprotein LprP	YES	1074754 Rev-Beg Scaf 35 1074756 Rev-End Scaf 13	1074754 Fwd-End Scaf 6 1074756 Rev-End Scaf 52	1074756 (CELERA Only)	YES	1074756 Fwd-Beg Scaf 18 GCC 1074754 Rev-Beg Scaf 17 GGC 1074754 Fwd-End Scaf 25 GCC	1074756 Fwd-Beg Scaf 12	1074756
3	H37Rv-emb|AL123456.3|	YES	YES	Rv1135c	PPE family protein PPE16	YES	1262961 Rev-Beg Scaf 27 1262963 Rev-End Scaf 16	1262966 Rev-Beg Scaf 54 1262963 Rev-End Scaf 40	1262963 (CELERA Only)	YES	1262961 Fwd-End Scaf 28 AGC 1262961 Fwd-End Scaf 99 1263062 Rev-Beg Scaf 20, 44 1262963 Rev-End Scaf 17 (-10bp mismatch)	1262961 Fwd-End Scaf 19, 56 1262961 Rev-Beg Scaf 96 GGCT 1262963 Rev-End Scaf 81 GGC (-10bp match at beginning of search)	1262963 REV_Beg/ END in CELERA Also
4	H37Rv-emb|AL123456.3|	YES	YES	Rv1371	Probable conserved membrane protein	YES	1543969 Fwd-End Scaf 40 1543972 Rev-End Scaf 37	1543972 Fwd-Beg Scaf 36 1543969 Fwd-End Scaf 63	1543972	YES	1543972 Fwd-Beg Scaf 27 AGG 1543969 Fwd-End Scaf 12 GAG	1543972 Fwd-Beg Scaf 89 1543969 Fwd-End Scaf 71	1543972
5	H37Rv-	YES	YES	Rv1469	Probable	YES	1657016	1657015 Fwd-	1657014	YES	1657015	1657015	1657014

	emb|AL123456.3|				cation transporter P-type ATPase D CtpD		Rev-End Scaf 46	Beg Scaf 27, 63 1657013 Rev-Beg Scaf 17			Fwd-Beg Scaf 12 CGT 1657013 Fwd-End Scaf 23 CGT	Fwd-Beg Scaf 71 1657013 Fwd-End Scaf 7 1657016 Rev-End Scaf 1 GGCG	
6	H37Rv-emb|AL123456.3|	YES	YES	Rv1522c	Probable conserved transmembrane transport protein MmpL12	YES	1715591 Fwd-End Scaf 3 1715599 Rev-End Scaf 42	1715591 Rev-Beg Scaf 46	1715593 (CELERA Only)	YES	1715593 Fwd-Beg Scaf 23 GTC 1715591 Fwd-End Scaf 6 GTC	1715593 Fwd-Beg Scaf 7	1715593
7	H37Rv-emb|AL123456.3|	YES	YES	Rv1754c	Conserved protein	YES	1986638 Fwd-Beg Scaf 60	1986638 Rev-End Scaf 55	1986637	YES	1986638 Fwd-Beg Scaf 6 TCT	1986638 Rev-End Scaf 80 GGCG 69 (-10bp match at beginning of search)	1986637
8	H37Rv-emb|AL123456.3|	YES	YES	Rv1917c	PPE family protein PPE34	NO (MIRA) YES (CELERA Only)	2165961 Fwd-Beg Scaf 71	2165961 Fwd-Beg Scaf 8 2165961 Rev-Beg Scaf 51	None (CELERA Only)	YES	2165921-2165956; 2165903-2165938 Fwd-End Scarf 173 2165903- 2165936; 2165920- 2165961 Rev-Beg Scaf 2 Rev-End 2165961-2165907 (43bp section missing) Scaf 14	2165961 Fwd-Beg Scaf 37 2165920 Fwd-End Scaf 96 2165961 Rev-End Scaf 83 (-43bp match at beginning of search) GGC	2165961 (CELERA Only)
9	H37Rv-emb|AL123456.3|	YES	YES	Rv2016	Hypothetical protein	YES	2263627 Fwd-Beg Scaf 44 2263625 Fwd-End Scaf 31	2263625 Rev-Beg Scaf 33 2263627 Rev-End Scaf 51	2263627 (CELERA Only)	YES	2263625 Fwd-End Scaf 20 TGA 2263627 Rev-End Scaf 2 (-10 bp match)	2263625 Rev-Beg Scaf 99 TCCT 2263627 Rev-End Scaf 13 GGC	2263627 (CELERA Only)
10	H37Rv-emb|AL123456.3|	YES	YES	Rv2435c	Probable cyclase (adenylyl-or guanylyl-)(adenylate-or guanylate-)	YES	2732276 Fwd-Beg Scaf 38 2732274 Rev-Beg Scaf 14	2732274 Rev-Beg Scaf 7	2732274 (CELERA Only)	YES	2732276 Fwd-Beg Scaf 29 AGC 2732274 Rev-Beg Scaf 22	2732276 Fwd-Beg Scaf 25 2732274 Fwd-End Scaf 75	2732276 (CELERA Only)
11	H37Rv-emb|AL123456.3|	YES	YES		MoaR1,transcriptional regulator	YES	3489916 Fwd-End Scaf 58 3489919		3489916 (CELERA Only)	YES	3489916 Fwd-End Scaf 33 CGG	3489918 Fwd-Beg Scaf 24	3489918 (CELERA Only)

					y protein		Rev-End Scaf 49				3489919 Rev-End Scaf 112 (-9bp mismatch)	3489916 Fwd-End Scaf 97	
12	CDC1551-gb|AE000516.2|	YES	YES	MT3429	Hypothetical Protein	YES	3708841 Fwd-Beg Scaf 65	3708841 Fwd-Beg Scaf 56	None	YES	3708841 Fwd-Beg Scaf 171 GTC 3708841 Rev-End Scaf 127	3708841 Rev-End Scaf 10 GGC 3708841 Rev-End Scaf 95 GGC	3708840
13	H37Rv-emb|AL123456.3|	YES	YES	Rv3128c	Conserved Hypothetical	YES (CELERA in FWD-END)	3493910 Fwd-Beg Scaf 58 3493911 Fwd-End Scaf 94 3493908 Rev-Beg Scaf 50 3493908 Rev-Beg Scaf 94	3493910 Fwd-Beg Scaf 9, 52 3493908 Fwd-End Scaf 9, 49	3493910	YES	3493910 Fwd-Beg Scaf 33 CGA 3493908 Fwd-End Scaf 44 CGA	3493907 Fwd-Beg Scaf 68, 97 3493908 Fwd-End Scaf 88 3493908 Rev-Beg Scaf 56 TCG	3493910 (CELERA Only)
14	H37Rv-emb|AL123456.3|	YES	YES	–	Intergenic Before Rv3180c	YES	3549199 Fwd-Beg Scaf 67,108 3549197 Fwd-End Scaf 48 3549199 Rev-End Scaf 108	3549199 Fwd-Beg Scaf 49 3549197 Fwd-End Scaf 56	3549199 (CELERA Only)	YES	3549199 Fwd-Beg Scaf 125 CGG 3549199 Scaf 127 CGGT 3549197 Rev-Beg Scaf 9, 3549197 Fwd-End Scaf 128,172 CGGA 3549223 Rev-Beg Scaf 172 (-32bp)	3549199 Fwd-Beg Scaf 88 3549197 Rev-Beg Scaf 95 TCGG Scaf 9 TCCG	3549199 (CELERA Only)
15	H37Rv-emb|AL123456.3|	YES	YES	Rv3427c	Possible transposase	YES	3844681 Fwd-Beg Scaf 79 3844678 Rev-Beg Scaf 59	3844678 Rev-Beg Scaf 48 Potential small rev-end Scaf 48 3844681 Rev-End Scaf 61	3844681	YES Indeterminate for CELERA	3844681 Fwd-Beg Scaf 87 GGG 3844678 Rev-Beg Scaf 41	3844678 Fwd-End Scaf 86 3844681 Rev-End Scaf 68,69 (-10bp match at beginning of search) GGCG	3844681 Indeterminate for CELERA
FWD_END: TGAACCACCTGACATGACCCCATCCTTTCCAAGAACTGGAGTCTCCGGACATGCCGGGGCGGTTCA													
16	H37Rv-emb|AL123456.3|	YES	YES	–	Intergenic Before Rv1765A	YES	1998623 Fwd-End Scaf 24	1998623 Fwd-End Scaf 10	1998623	YES	1998623 Rev-Beg Scaf 14	1998623 Rev-Beg Scaf 83 GGTG	1998811
REV_BEG: TGAACCGCCCCGGCATGTCCGGAGACTCCAGTTCTTGGAAAGGATGGGGTCATGTCA													
17	H37Rv-	YES	YES	–	Intergeni	YES	888992 Rev-	888978 Fwd-End	888992	YES	888992	888990	888992

	emb|AL123456.3|				c Before Rv0795		Beg Scaf 22 Rev-beg Scaf 103	Scaf 36 888992 Fwd-End Scaf 61 888787 Rev Begin Scaf 54			Fwd-End Scaf 9 TGG 888787 Fwd-End Scaf 179	Fwd-Beg Scaf 34, 35 888992 Fwd-End Scaf 85 888787 Fwd-End Scaf 57,58,59 888787 Rev-Beg Scaf 57 CGG	(CELERA Only)
18	H37Rv-emb|AL123456.3|	YES	YES	–	Intergenic Before Rv0002	YES	1889 Fwd-Beg Scaf 11 1892 Fwd-End Scaf 72	1892 Rev-Beg Scaf 29 1886 Rev-End Scaf 4	1892	NO			None
19	H37Rv-emb|AL123456.3|	YES	YES	–	Intergenic Before PE22 (Rv2107)	YES	2366894 Fwd-End Scaf 32 2366890 Rev-End Scaf 26	2366894 Fwd-End Scaf 60 2366890 Rev-End Scaf 18	2366894 (CELERA in REV-END)	YES	2366894 Rev-Beg Scaf 20 2366890 Rev-End Scaf 13	2366894 Fwd-End Scaf 99	2366894 (CELERA Only) (CELERA in REV-END)
20	H37Rv-emb|AL123456.3|	YES	YES	Rv2352c	PPE family protein PPE38	YES	2634022 Fwd-Beg Scaf 36 2634048 Fwd-End Scaf 33	2634022 Fwd-Beg Scaf 7 2634048 Rev-Beg Scaf 18	2634048 (CELERA in REV-END	YES	2634048 Fwd-End Scaf 11 CTA 2634022 Rev-End Scaf 29 (-10bp mismatch)	2634022 Fwd-Beg Scaf 55 2634048 Fwd-End Scaf 2, 74	2634048 (CELERA in REV-END
21	H37Rv-emb|AL123456.3|	YES	YES	–	Intergenic Beore Rv2814c	YES	3127923 Rev-End Scaf 20	3127932 Fwd-Beg Scaf 50 3120468 Rev-Beg Scaf 47	3119350 (M) 3122552 (C) (RD207)	YES	3120468 Fwd-End Scaf 26 3127923 Rev-End Scaf 4	3127923 Rev-End Scaf 82 CGG (-2bp match at beginning of search)	3119200 (RD207) 3120248
22	H37Rv-emb|AL123456.3|	YES	NO	–	Intergenic Before Rv3019c	YES	3378553 Fwd-End Scaf 86		3378553	YES	3378553 Rev-Beg Scaf 184	3378553 Rev-Beg Scaf 66,82 TGGC	3378553
23	H37Rv-emb|AL123456.3|	YES	YES	Rv3383c	Possible polyprenyl synthetase IdsB (polyprenyl transferase) (polyprenyl diphosphate synthase)	YES	3797827 Fwd-End Scaf 54 3797823 Rev-End Scaf 51 3797823 Rev-End Scaf 115	3797827 Fwd-End Scaf 57 3797823 Rev-End Scaf 11	3797827 (CELERA in REV-END)	YES	3797827 Fwd-End Scaf 94 GAT 3797823 Rev-End Scaf 41, 78, 183 (-10bp mismatch)	3797825 Fwd-Beg Scaf 69 3797826 Fwd-End Scaf 15 3797827 Fwd-End Scaf 47	3797827

24	H37Rv-emb|AL123456.3|	YES	YES	–	Intergenic After Rv0001	YES	1592 Fwd-Beg Scasf 72, 73 Fwd-End Scaf 23	1594 Fwd-End Scaf 16 1592 Rev-End Scaf 29	1594	YES	1594 Rev-Beg Scaf 24 1592 Rev-End Scaf 30 (-10bp mismatch)	1592 Fwd-Beg Scaf 3 1594 Rev-Beg Scaf 98 GATT	1594
25	Mycobacterium bovis subsp. bovis AF2122/97 complete genome			Intergenic Before Mb2837 (Rv2813)	Mb2837 REQUIRED FOR THE TRANSPOSITION OF THE INSERTION ELEMENT IS6110							3076124 Rev-Beg Scaf 50 TCTC	
						REV_END: TCAACGCCAGAGACCAGCCGCCGGCTGAGGTCTCAGATCAGAGAGTCTCCGGACTCACCGG						
26	H37Rv-emb|AL123456.3|			–	In Intergenic region before **Rv3326** - Putative transposase for insertion sequence element IS6110	YES (MIRA) NO (CELERA)			3711737	No			None
27		YES	YES	Rv3020c	ESAT-6 like protein EsxS	No (MIRA) YES (CELERA)	3379025 Fwd-Beg Scaf 80	3379025 Fwd-Beg Scaf 14	None 3379025	YES		3379025 Fwd-Beg Scaf 71	3379025 FWD_Beg in CELERA

3.27 MIRU/VNTR repeats

There were a total of 14 tandem repeats identified in each of the assembled genomes which were not common between the 2 genomes as shown in the Tables 3.27 and 3.28. These did not match any of the MIRU/VNTR repeats used for molecular epidemiology proposed by Supply *et al*., (2006). A manual search in the genome assemblies revealed a single repeat for each locus under 24 loci MIRU/VNTR typing suggesting that the sequence assembly cannot resolve the MIRU/VNTR repeats.

Table 3.27: Tandem repeats identified using JEMBOS in the hypo-virulent strain with 100% identity match

Start	End	Size	Count	Identity	Consensus
3104690	3105046	51	7	100	tgacctccgccggcgacgatgcagagcgcagcgatgaggaggagcggcgct
1956264	1956491	57	4	100	ctagcgtggcgacgatgcgggctgggatgggccgctgaggagccgcgcggtcgagct
3445858	3446031	58	3	100	agtgccgatcgcaagcgcggcgctagcgccgggcgcggcgggtcggcaccatcgggct
3099216	3099380	55	3	100	agcggtgatcgcaagcgcggcgagccgggcgcagcgggtcaccgccatcgggact
1665750	1665908	53	3	100	ggcgcggttcacgagcgccgctcctcctcatcgcttcgctctgcatcgtcgtc
2122397	2122550	77	2	100	gagcggcgccaatgagccgcgccggcgacgatgcagtgggggtaccgcccgcttgcgggggacgaagcgatgacgag
3777128	3777253	63	2	100	cctgtgagtcgagtgagcggaacgaacgaagtgagtgacgggaacgagacgaacaatccggcc
3445682	3445799	59	2	100	agtgccgatcgcaagcgcggcgcttgcgccgggcgcggcgggtcggcaccatcgggcta
4275081	4275194	57	2	100	ccgtgacgatcgcgagcccggcgcagccgggcgaagcgggtcggcacgcatcggacc
2541242	2541355	57	2	100	agcccggcgacgacccgggtcggcacgacccgggaaggaaccgggcaaatcaagcac
3460384	3460495	56	2	100	ccctgccgggcgacgattcgggccggcacggcccgatgaggagcccggcaatcaga
1965885	1965996	56	2	100	agtcgggtgacgatgcgggccggtgtggtccgaggaggagcccgacaatttaagct
1462071	1462182	56	2	100	ggttagattgcccggctcctcaacccgccgtttcggcgtgcatcgtcgccgggcta

Table 3.28*:* Tandem repeats identified using JEMBOS in the hyper-virulent strain with 100% identity match

Start	End	Size	Count	Identity	Consensus
3042339	3042593	51	5	100	tgacctccgccggcgacgatgcagagcgcagcgatgaggaggagcggcgct
1911119	1911346	57	4	100	ctagcgtggcgacgatgcgggctgggatgggccgctgaggagccgcgcggtcgagct
3383891	3384064	58	3	100	agtgccgatcgcaagcgcggcgctagcgccgggcgcggcgggtcggcaccatcgggct
3036865	3037029	55	3	100	agcggtgatcgcaagcgcggcgagccgggcgcagcgggtcaccgccatcgggact
1656593	1656751	53	3	100	ggcgcggttcacgagcgccgctcctcctcatcgcttcgctctgcatcgtcgtc
2075657	2075810	77	2	100	gagcggcgccaatgagccgcgccggcgacgatgcagtgggggtaccgcccgcttgcgggggacgaagcgatgacgag
3714533	3714658	63	2	100	cctgtgagtcgagtgagcggaacgaacgaagtgagtgacgggaacgagacgaacaatccggcc
3383715	3383832	59	2	100	agtgccgatcgcaagcgcggcgcttgcgccgggcgcggcgggtcggcaccatcgggcta
4216003	4216116	57	2	100	ccgtgacgatcgcgagcccggcgcagccgggcgaagcgggtcggcacgcatcggacc
2501326	2501439	57	2	100	agcccggcgacgacccgggtcggcacgacccgggaaggaaccgggcaaatcaagcac
4339350	4339461	56	2	100	ctagcccggcgacgatgcacgccgaaacggcgggttgaggagccgggcaatctaac
3398417	3398528	56	2	100	ccctgccgggcgacgattcgggccggcacggcccgatgaggagcccggcaatcaga
1920740	1920851	56	2	100	agtcgggtgacgatgcgggccggtgtggtccgaggaggagcccgacaatttaagct

4 DISCUSSION

We carried out a comparative whole genome sequence analysis of *Mycobacterium tuberculosis* strains representative of the previously described seven sub-lineages of the *M. tuberculosis* Beijing lineage found in Cape Town, South Africa. In addition, a more focused genome analysis was also done on two members of sub-lineage 7 which had previously been reported to have contrasting phenotypes with respect to disease transmission and their ability to kill in the mouse model (Aguilar *et al*., 2010). All strains used in the study had their genomes sequenced on the Illumina Highseq2000 platform to generate paired-end sequence data. Sequence data was subsequently mapped to the *M. tuberculosis* H37Rv reference genome using 3 mapping algorithms. Mapping statistics of the different software illustrated that there were slight differences in how the mapping algorithms aligned reads to the reference. To this end, there were associated differences in SNP calls made relative to the reference of all genomes. In order to increase the confidence of SNPs called for aligned genomes, an overlap of SNPs identified in 3 alignments were used to determine high confidence SNPs. The high concordance found when compared to 273 SNPs that were previously verified in 2 of the genomes in this study suggested that this approach for SNP calling had 99.6% accuracy.

This study included a comparative analysis of phylogenetic trees generated based on genome-wide high confidence SNPs (current study) and previously constructed phylogenetic trees based on replication, repair and recombination (3R) genes (Mestre *et al*., 2011) and one based on insertion points and selected SNPs in the *ogt* and *mutT* genes (Hanekom *et al*., 2007) was done. There was a significant difference between the phylogeny based on 3R genes described by Mestre *et al.* (2011) and that based on genome-wide SNPs in this study. Branch collapse was noticed in the phylogenetic tree based on the 3R genes of Mestre *et al.* (2011) compared to the high resolution of the tree based on genome-wide SNPs. Possible reasons for the differences in resolution of the aforementioned phylogenetic trees could be attributed to the use of a global representation of strains in the study by Mestre *et al.* (2011) compared to strains obtained in a single, high tuberculosis incidence endemic area in the present study. Furthermore, the addition of genomes from the study of Schürch *et al.* (2010) did not affect the resolution of genome-wide phylogenetic trees in the current study. The comparison of phylogenetic trees based

on the study by Hanekom *et al*. (2007) and the ones in this study based on genome-wide SNPs revealed that sub-lineage 5 and 6 interchanged positions. Sub-lineage 6 was subsequently determined to be evolutionarily older than sub-lineage 5 in the present study in contrast to the research published by Hanekom *et al.* (2007) where the opposite was found. This difference was attributed to the use of more SNP positions in the present genome-wide study compared to that done by Hanekom *et al.* (2007). Additionally, sub-lineages 2 and 3 were found to be more closely related than previously reported in by Hanekom *et al.* (2007).

The genomic markers used by Hanekom *et al* (2007) to delineate sub-lineages 2,3 and 4 as well as sublineages 5 and 6 proved to have better resolution compared to the selected Beijing SNP type markers used by Schurch *et al* (2011). This is on account of the failure to determine the evolutionary order of sub-lineages 5 and 6 under Schurch *et al* (2011) due to them sharing the same Beijing SNP Type marker. Our study provides some evidence for a revision of the evolutionary order of the 7 sub-lineages of Beijing in Cape Town based on genome wide SNPs. It has been observed in this study that when SNPs found in the categories of PE/PPE and insertion sequences and phages are removed from analysis, sub-lineage 5 has 1 unique intergenic SNP shared with sub-lineage 7 where as no unique SNPs are shared between sub-lineages 6 and 7. Additionally, sub-lineages 2 and 3 are very closely related than previously reported when considering genome-wide SNPs. There are 28 SNP differences between the two when excluding SNPs in the categories of PE/PPE and insertion sequences and phages which is comparable with the level of SNP differences between the hypo-hyper virulent strains of sub-lineage 7 which is suggestive that they belong to the same sub-lineage. This is in agreement with the phylogenetic tree generated in this study and is suggestive of there being 6 sub-lineages of Beijing in the Cape Town when considering genome-wide SNPs. Additionally sub-lineage 1 appears to be an out-group to the 6 other sub-lineages to which they share a common ancestor when looking at the phylogenetic tree in this study. This supports the scenario that different sub-lineages might have been imported into Cape Town at different time points and seeded and evolved independently with differing success (Hanekom *et al*., 2007a; Luo *et al*., 2015; Merker *et al*., 2015). Taking into account the aforementioned evolutionary scenario in this study, the inclusion of a much larger sample size incorporating previously described sub-lineages 2,3,5 and 6 strains would increase our confidence for a

revised phylogeny. Estimating the time since sequences diverged can be done and relies on molecular clocks (Brites and Gagneux, 2015).This has been utilised to estimate the divergence of *M.*tuberculosis in a numberof studies including that of Merker et al., (2015) and Comas et al., (2013) and relies on a number of models and assumptions. This includes the co-evolution of *M.tuberculosis* with its human host when calculating age and time of divergence of the *M.tuberculosis* (Brites and Gagneux, 2015; Comas *et a*l., 2013; Merker *et al*., 2015). Looking at the evolution and divergent times of the Beijing lineage in South Africa, Brites and Gagneux, (2015) did suggest that a co-evolution of the Beijing lineage with its human host is unlikely to be inferred considering the relative recent introduction of the lineage to South Africa. However, investigations using histological samples reported by Cowley *et al*., (2008) dating back to the 1930s-2005 could potentially be used for calibration of molecular clocks for Beijing lineage evolution in Cape Town albeit strains being isolated from children only. However, the genetic markers used by Hanekom *et al*. (2007) were able to delineate sub-lineages 2 and 3 better than genome-wide SNP markers and could thus have an impact on clinical decision making with regards to a strain described as being part of a transmission chain, a unique event or reactivation of a previous infection.

In order to understand the evolution of the Beijing lineage in this study, a focused analysis of the evolution of SNPs leading to amino acid changes was undertaken. This analysis showed that evolution within each sub-lineage was mainly in the functional groupings of cell wall and cell processes, and intermediary metabolism and respiration. Sub-lineage 1, as described by (Hanekom *et al*., 2007a), was shown by phylogenetic analysis in the present study to be the most ancestral of the Beijing strains in our geographical area. Molecular epidemiology data have previously shown that this sub-lineage does not spread well in our study area and is less virulent in the mouse model (Aguilar *et al*., 2010). The Beijing sub-lineage 5 has been shown to have low transmission ability of its drug sensitive strains and a high capability of its multi-drug resistant strains to transmit well. Previous studies had shown that sub-lineage 6 was a moderately transmitted strain in the Cape Town area (Hanekom *et al*., 2007a). The Beijing sub-lineage 7 has been reported to be responsible for the majority of tuberculosis cases in Cape Town (Hanekom *et al*.,

2007a; van der Spuy *et al*., 2009). Furthermore, this lineage had been shown to be highly virulent and transmittable in a mouse model (Aguilar *et al.,* 2010). Following on from the description of evolutionary changes that were unique to each sub-lineage, the common ancestors at branch points of phylogenetic trees had similar functional groups in which amino acid changes occurred at sub-lineage level the branch node encompassing sub-lineages 2, 3, 4, 5, 6 and 7. These had intermediary metabolism and respiration as the major functional group harboring SNPs. This is suggestive that amino acid changes in the evolution of the Beijing lineage in this study occurs mainly in the functional grouping of intermediary metabolism and respiration, cell wall and cell processes, conserved hypothetical proteins and regulatory proteins. Subsequent analysis of sub-lineage defining nsSNPs and phylogenetic tree branch node common ancestor nsSNPs for overrepresentation of biological processes using the PANTHER classification system revealed results for unique nsSNPs for sub-lineage 1 and combined sub-lineages 2 and 3 as well as for the common ancestor for branch node common ancestor for sub-lineages 2, 3, 4, 5, 6 and 7. Taking into consideration the results of the unique genomic sub-lineage nsSNPs and the overrepresentation results of the PANTHER biological processes, evolution within the 7 sub-lineages of Beijing in this study occurred to a great extent in the atypical sub-lineages encompassing sub-lineages 1, 2 and 3. More focused investigations involving expression data like whole proteome analyses are needed to further elucidate the effects of the aforementioned nsSNP differences within the 7 sub-lineages of Beijing in this study. The overrepresentation of nsSNPs in genes according to the PANTHER algorithm with respect to the 7 sub-lineages in this study was suggestive of evolution taking place in a similar manner across the lineages. This, however, needs to be viewed in the context that very few strains were included in this study which could have affected our results. Focussing on the groupings of typical and atypical Beijing strains, the results of this study were different from those shown by (Schürch et al., 2011b) in the sense that the regulatory functional grouping for nsSNPs were not overrepresented in the typical Beijing strain representatives. Genes involved in the functional groupings of lipid metabolism had the highest number of nsSNPs followed by those involved in intermediary metabolism and cell wall and cell processes groups. Related to this is the higher proportion of informative nsSNPs in the lipid metabolism functional grouping for lineage 2 (encompasses Beijing) when compared to other lineages Coll *et al*., (2014). The ex-vivo study of

Mendum *et al*., 2015 highlighted the consequences of having functional mutations in these groupings and the significant relative numbers of nsSNPs in this grouping could play a role in the higher success of typical Beijing *M. tuberculosis* strains compared to atypical Beijing strains in Cape Town (Hanekom *et al*., 2007b; Mendum *et al*., 2015; van der Spuy *et al*., 2009). However, this needs to be further investigated using more strains in the analysis to determine whether the results in this study were as a result of not having a large enough sample size and also bearing in mind that many more factors might play a role in the success of *M. tuberculosis*. This includes the fact that even though nsSNPs in the regulatory functional groupings are not overrepresented, functional deleterious SNPs in this group can have a pronounced effect as has been demonstratedted in H37Rva (Lee *et al*., 2008). Furthermore, other geographical areas have shown atypical Beijing strains being more successful than the typical modern strains as is the case in Japan and the Eastern Cape of South Africa albeit only with respect to drug resistant strain in the latter (Klopper *et al*., 2013; Millet *et al*., 2012; Wada *et al*., 2009).

Furthermore the identification of unique intergenic SNPs in close proximity of previously reported TSS warrants additional investigations of the possible effects of variants in the promoter regions of genes. Studies have been carried out on the influence of genomic changes taking place in the promoter regions of genes (Bashyam and Tyagi, 1998; Cortes *et al*., 2013; Newton-Foot and Gey van Pittius, 2013). Changes within 3 base pairs of the -10 promoter region have been reported to have an effect on transcription. Investigations in this study did find such genomic changes in *Rv1215c* in sub-lineage 5 lineage. The protein of this gene has been classified as a conserved hypothetical protein of unknown function under Tuberculist and identified in the membrane fractions of proteomics investigations (Malen et al., 2011; de Souza and Wiker, 2011). On account of this and that sub-lineage 5 has been reported to successfully transmit drug resistance strains, one could hypothesise that mutations in this region could play a role in this exhibited phenotype and thus should be further investigated. Furthermore, with a reported figure of 24% transcription binding sites located in intergenic regions, more research utilising whole genome sequencing and expression studies can guide and help elucidate role of intergenic SNPs in regulation in *M. tuberculosis* (Newton-Foot and Gey van Pittius, 2013; Cortes *et al*., 2013; Galagan *et al*., 2013; Minch *et al*., 2015).

The availability of molecular epidemiology and proteomics expression data and mouse model infection comparative studies for 2 members of sub-lineage 7 highlight the importance of a holistic approach when trying to elucidate the effects of genomic evolution in *M. tuberculosis*. In spite of the high similarity in their genomes, the hypo-virulent strain (SAWC 507) and hyper-virulent strain (SAWC 5527) of Beijing sub-lineage 7 had contrasting phenotypes in their ability to transmit, immune response and ability to kill in the mouse model. The functional annotation of their SNPs with respect to gene products using 4G SIFT and subsequent comparison to protein expression data illustrated how confidence can be increased in inferring consequences of genomic changes. Agreement in this regard was noted in 3 functionally deleterious mutations identified by 4G SIFT and accompanied by lack of expression in the proteomics data. Rv0988 is an exported protein of unknown function (Lew *et al*., 2011; de Souza *et al*., 2011). The hypo-virulent strain had a deleterious mutation for this gene whilst the hyper-virulent strain didn't. The proteomics data didn't show expression for both strains probably attributed to the absence of culture filtrate proteomics investigations. This is in line with the work of de Souza *et al*.(2011) where this protein was absent in the whole cell lysate but present in the culture filtrate of H37Rv. On the other hand, Rv0214 which is associated with lipid metabolism had a deleterious nsSNP in the hyper-virulent strain but not in the hypo-virulent strain. This was supported by proteomics data and has previously been identified in lungs of infected guinea pigs (Kruh *et al*., 2010). However, Rv0214 is not essential for in vivo survival of the pathogen which is in line with the mouse model studies of the hypo-hyper virulent strains (Aguilar *et al*., 2010; Griffin *et al*., 2011). In this study we found that the product for *Rv3542c* was not present in the hyper-virulent strain proteomics data which is in agreement with the SIFT prediction algorithm for deleterious functional mutations. Rv3542c has been shown to be involved in the importation of cholesterol which is important in chronic infection and survival in macrophages (Chang *et al*., 2009; Rengarajan *et al*., 2005). This, however, is in contrast to the phenotype of the hyper-virulent strain in its ability to kill in the mouse model study of Aguilar *et al*., 2010 and associated ability to cause disease from an epidemiological perspective. We thus hypothesize that some compensatory genomic changes for this for this deleterious mutation with respect to survival in the macrophage has taken place and warrants further investigation. Looking at Rv2493 which was identified in proteomics investigation for the hyper-

virulent strain, the study by Kruh *et al*., 2010 showed it to be present in the lungs of *M. tuberculosis* infected guinea pigs. The hypo-virulent strain was predicted to have a functional deleterious mutation for this gene by SIFT and had no expressed protein in the investigation of de Souza *et al*.,(2010). Mendum *et al*., (2015) demonstrated that a functional loss of Rv2493 has a fitness cost for *M. tuberculosis* in dendritic cells as well as in macrophages and mice. This is likely to be associated with failure to prevent phagosomal acidification which is required for *M. tuberculosis* killing in a host (Mendum *et al*., 2015). Taken together this is suggestive that *Rv2934* could be involved in virulence within the context of this study.

The comparative analysis of small indels and whole cell lysate protein expression in the hypo-hyper virulent strains revealed that there was no direct influence of small indels on the latter in an in-vitro model. It however remains to be elucidated whether there is an influence on the expression of culture filterate proteins which could have an effect on the interation of the pathogen with its host. The large duplication events analysis 2 strains did show that 2 potential large duplications greater than 500bp exhibited by sequence coverage greater than 1.8x that of the mean coverage. When this was compared to the results of Weiner *et al*., (2012) the large duplication events in this study differed. In addition, the large duplication event identified by Domenech *et al*., (2010) was also not identified. Verification of these duplications is however yet to be done.

The analysis of tandem repeats in the assembled genome using JEMBOS showed 1 variation in the number of tandem repeats found between the hypo and hyper assembled genomes. However, MIRU/VNTR repeats used under the standerdised 24 loci set proposed by Supply *et al*., (2006) could not be picked up by JEMBOS. Manual inspection of the number of MIRU/VNTR repeats in each assembly showed that there was only one copy of each MIRU/VNTR. This explains why JEMBOS was not able to pick up any of the above mentioned MIRU/ VNTR repeats because it reports cases where there are at least 2 or more copies of a repeat. The repeat unit that showed variation in copy number between the hypo-hyper strains as determined by JEMBOS was identified to be a MIRU class II element. The gene adjacent to it, Rv2680, had equivalent proteomic expression in the hypo-hyper strains suggestive that there was no direct influence due to their difference in repeat copy number. Both

MIRA and CELERA had difficulty resolving more than 1 repeat of Supply *et al*., (2006) MIRU/VNTR loci taking into account the results of Hanekom *et al*.(2007c) which showed that members of sub-lineage 7 had more than 1 repeat in 11 out of 12 MIRU loci. This however would still need to be confirmed by wet bench investigations. We hypothesize that with the use of both paired and mate pair sequences in genome assembly, such difficulties will be resolved.

Structural variation in the genome of *M. tuberculosis* can also contribute to its evolution through the number and positioning of the transposon element IS*6110*. Determining the exact location of the IS*6110* from whole genome assembly of NGS data can however be a challenge as a result of the repetitive nature of the transposon element and the whole *M. tuberculosis* genome in general. This was also limited in this study by the Illumina paired end short reads having a read length of 105bp (shorter than the repeat element searched for). An association between the location of the IS*6110* element and a previously reported genomic inversion event was identified in the *de novo* assembly of the hypo-virulent (SAWC 507) strain. The inversion event was, however, only detected when using one of the two genome assembly software tools. *De novo* assembly was able to identify structural variation between the hypo- and hyper-virulent strains of sub-lineage 7 through IS*6110* location and potential genomic rearrangement as exemplified by genomic inversions. This adds a level of genomic evolution beyond SNP variation that can potentially affect virulence phenotype. In the absence of mate pair sequences with large insert sizes, the confidence of the *de novo* assemblies were not high, despite having long read PacBio sequences. Investigations by others have indicated that assembly accuracies can be compromised if coverage of the PacBio is less than 80x and with no accompanying sequences with a large insert size (e.g. mate pairs) (Chin *et al*., 2013; Koren and Phillippy, 2015; Koren *et al*., 2012). This can potentially affect the resolution of number of gaps in assembly needing closure, large repetitive areas in a genome and downstream analysis like the location of the transposon element IS*6110*. In spite of the aforementioned limitations, agreement was observed for the identification of the element IS*6110* using 2 different assemblers at over 90% of positions in 2 strains compared. Furthermore, the IS*6110* element located at position 1892 and unique to the hypo-virulent strain was located upstream of Rv0002. However, the transcriptional start site and putative promoter elements as shown by

Chauhan and Tyagi,(2011) were downstream of this region and thus insertion was not directly indicative of having an effect on the transcription of Rv0002. Furthermore Rv0002 protein was not differentially expressed when looking at the results of de Souza *et al*., (2010) for the hypo and hyper-virulent strains.

.

The holistic approach of using WGS analysis with expression has highlighted the value of this and the particular roles which Rv0214 and Rv2493 in particular may play in understanding the underlying reasons explaining the phenotypic differences between 2 closely related strains. Furthermore, the use of SIFT as a predictor of deleterious mutations was validated in the strains with proteomics expression data and can thus be a useful to for directing investigations for which whole genome sequencing data exists. However, structural variation investigations especially when considering large genomic repeats was hampered by the lack of sequences with large insert sizes.

5 CONCLUSION

The current study described and analysed the previously described 7 sub-lineages of the Beijing lineage of *M. tuberculosis* in Cape Town South Africa using whole genome sequencing. Three mapping algorithms were used to align sequence reads of representatives of the 7 sub-lineages of Beijing to the *M. tuberculosis* H37Rv reference genome and a set of high confidence SNPs were derived by using an overlap of SNPs derived from the alignments produced by the 3 mapping algorithms. The sequencing for this purpose was done on an Illumina HiSeq2000 platform and yielded 105bp paired-end sequencing reads. The accuracy of using the overlap of the 3 mapping algorithms to determine high confidence SNPs was verified at 273 verified SNP positions. This approach was determined to be 99.6% accurate. Additionally, 2 strains of sub-lineage 7 which were previously reported to be hypo- ad hyper-virulent with accompanying proteomics expression data and mouse model infection data were sequenced on the Pacbio platform. Sequence reads from the Pacbio platform were used together with the Illumina reads to perform hybrid *de novo* assembly.

The high confidence SNPs were used to construct phylogenetic trees using maximum likelihood, parsimony and neighbour-joining algorithms with 1000 bootstrap replicates. These were all congruent with a patristic distance correlation of 1. Subsequent comparative analysis using genome-wide SNPs identified in this study and molecular genetic markers used by others (Hanekom *et al.*, 2007) for molecular epidemiology showed that sub-lineages 5 and 6 were interchanged in terms of order of evolution in the phylogenetic tree. Additionally, the markers used by Hanekom *et al.* (2007) had a higher resolution in distinguishing sub-lineages 2 and 3 compared to the present study. The comparison to the phylogenetic tree based on the 3R genes (Mestre *et al.,* 2011) showed that it was not 100% congruent with the genome-wide SNP-based phylogenetic trees produced here, and had a patristic correlation of 0.8034. Taken together, the evolutionary scenario of the 7 sub-lineages in this study were best resolved by the genome-wide SNPs when compared to the evolutionary markers used by Hanekom *et al.* (2007) and Mestre *et al.* (2011). However the markers used by Hanekom *et al.* (2007) were better suited for molecular epidemiologic studies on account of better resolving sub-lineages 2 and 3 compared to the present study. Additionally, sub-lineage 1 appears to be an out-

group of the other 6 sub-lineages based on the topology of genome-wide SNP-based phylogenetic trees in the current study. This is suggestive of strains being brought into Cape Town at different time points after which they got established and consequently evolved with different success rates. Sub-lineage 2 and -3 were also observed to be very closely related and probably arose from the same common ancestor in the near past.

The analysis of the amino acid change evolution on the 7 sub-lineages of Beijing based on the genome wide SNPs indicated that evolution within individual sub-lineages was generally occurring in similar functional groupings of cell wall and cell processes, and intermediary metabolism and respiration as defined in TubercuList. The same was also true when considering common ancestors represented at the branch node of phylogenetic trees. However, an analysis of overrepresentation of biological processes using the PANTHER analysis pathway showed that only sub-lineage 1 and sub-lineages 2 and -3 combined had enrichment of SNPs in genes involved in particular biological processes. These were the primary metabolic processes and metabolic process for sub-lineage 1 and protein acetylation, cellular protein modification processes and protein metabolic processes for sub-lineage 2 and 3. Additionally, unique intergenic SNPs identified in close proximity of previously described transcriptional start sites could have a possible effect on gene expression contributing the transmissibility or virulence of a strain. This is suggestive of the clonal evolution of *M. tuberculosis* and selection of certain traits which results in the creation of unique sub-lineages. The success of the sub-lineages is subsequently determined by the genomic traits it accumulates and the continuous evolution it undergoes.

The focussed analysis of 2 closely related members of Beijing sub-lineage 7 that had been previously shown to exhibit contrasting virulence phenotypes revealed that they had unique functional SNPs as determined 4G SIFT annotation.

In conclusion, this study highlights the importance and resolution provided by the use whole genome sequencing data to better understand the biology and evolution of *M. tuberculosis.* The use of protein expression data in conjunction with whole genome sequencing further enhanced this understanding by reconciling genomic and proteomics data. Structural variation when utilising whole genome sequencing was also appreciated in terms of the possible effects of the transposon IS*6110* and

associated large sequence inversion event.

6 KNOWLEDGE GAPS AND FUTURE STUDIES

Sub-lineages within a strain family can exhibit different phenotypes. Considering the Beijing Family, different sub-lineages have exhibited variation in both transmission patterns as well as animal model studies showing how dissimilar strains can be in terms of causing fatality (Aguilar *et al*., 2010). Taking into consideration such studies that look only at broadly defined lineages, rather than sub-lineages when describing phenotypes, conclusions drawn from such works can be spurious. The Beijing sub-lineage 7 has been the best characterised lineage in this respect and highlights that clear differences can exist even within the sub-lineages. What remains to be done are similar characterizations of other sub-lineages of Beijing and other PGG1 members to the same extent as the Beijing sub-lineage 7. A more robust evolutionary scenario can subsequently be established and when viewed in parallel with observed phenotypes, it can be used to better understand virulence and regulation in the Beijing sub-lineages and in general of the PGG1 members of *M. tuberculosis*. Being able to describe what parts of the genome have evolved to play a role in establishing the observed phenotype can subsequently be useful in identifying drug targets as well as identification of vaccine candidates (Kumar *et al*., 2010). Considering vaccine development, it has been reported in a number of studies that the current vaccine, BCG (Bacillus Calmette–Guérin), has had low efficacy in areas endemic to PGG1 members like in South India (Narayanan, 2006). Investigations focusing on gene products involved in modulating the immune system have provided candidates that can be used in the generation of vaccine candidates (Beaulieu *et al*., 2010). Closely related to this are candidates having no sequence homology to the human genome and are thus more likely candidates as drug targets. Targets identified in the study by Beaulieu *et al.* (2010) can also be looked at in proteomic studies covering whole cell lysis, secreted as well as membrane bound proteins. Those that are over- or under expressed can be further investigated by looking at the immune response elicited by the infecting strain and subsequently comparing this to other observed phenotypes, such as the ability to kill and epidemiological transmission data in specific human populations. Thus, on account of the high proportion of TB cases globally attributed to the Beijing lineage and PGG1 members in general, a high resolution characterization of strain members could ultimately contribute to the control of global TB.

Within the context of the current study, a number of future works can be carried out and include the following:

1. Proteomics study of 7 sub-lineages of Beijing in association with mouse model infection model studies.
2. Promoter investigation looking at the functional effect of intergenic SNPs in close proximity to previously described TSS.
3. Functional studies on SNPs that have been reconciled by proteomics and 4G SIFT annotation results.
4. Sequencing of strains using mate pair libraries with large insert sizes to enable better structural variation studies such as location of IS*6110*, inverted large sequences in relation to a reference strain and more accurate identification of large duplication events.

7 REFERENCES

Aabye, M.G., Ravn, P., Johansen, I.S., Eugen-Olsen, J., and Ruhwald, M. (2011). Incubation of whole blood at 39°C augments gamma interferon (IFN-γ)-induced protein 10 and IFN-γ responses to *Mycobacterium tuberculosis* antigens. Clin. Vaccine Immunol. *18*, 1150–1156.

Abadia, E., Zhang, J., Vultos, T. dos, Ritacco, V., Kremer, K., Aktas, E., Matsumoto, T., Refregier, G., Soolingen, D. van, Gicquel, B., et al. (2010). Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infection, Genetics and Evolution *10*, 1066–1074.

Aguilar, D., Hanekom, M., Mata, D., Gey van Pittius, N.C., van Helden, P.D., Warren, R.M., and Hernandez-Pando, R. (2010). *Mycobacterium tuberculosis* strains with the Beijing genotype demonstrate variability in virulence associated with transmission. Tuberculosis (Edinb) *90*, 319–325.

Alland, D., Lacher, D.W., Hazbón, M.H., Motiwala, A.S., Qi, W., Fleischmann, R.D., and Whittam, T.S. (2007). Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. J. Clin. Microbiol. *45*, 39–46.

Alonso, H., Samper, S., Martín, C., and Otal, I. (2013). Mapping IS6110 in high-copy number *Mycobacterium tuberculosis* strains shows specific insertion points in the Beijing genotype. BMC Genomics *14*, 422.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Arora, J., Singh, U.B., Suresh, N., Rana, T., Porwal, C., Kaushik, A., and Pande, J.N. (2009). Characterization of predominant *Mycobacterium tuberculosis* strains from different subpopulations of India. Infect. Genet. Evol *9*, 832–839.

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., and Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics *25*, 1968–1969.

Baker, M. (2012). De novo genome assembly: what every biologist should know. Nat Meth *9*, 333–337.

Baker, L., Brown, T., Maiden, M.C., and Drobniewski, F. (2004). Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis* . Emerging Infect. Dis. *10*, 1568–1577.

Bashyam, M.D., and Tyagi, A.K. (1998). Identification and analysis of “extended -10” promoters from mycobacteria. J. Bacteriol. *180*, 2568–2573.

Beamer, G.L., Flaherty, D.K., Assogba, B.D., Stromberg, P., Gonzalez-Juarrero, M., de Waal Malefyt, R., Vesosky, B., and Turner, J. (2008). Interleukin-10 Promotes

Mycobacterium tuberculosis Disease Progression in CBA/J Mice. J Immunol *181*, 5545–5550.

Beaulieu, A.M., Rath, P., Imhof, M., Siddall, M.E., Roberts, J., Schnappinger, D., and Nathan, C.F. (2010). Genome-Wide Screen for *Mycobacterium tuberculosis* Genes That Regulate Host Immunity. PLoS ONE *5*, e15120.

Belay, M., Ameni, G., Bjune, G., Couvin, D., Rastogi, N., and Abebe, F. (2014). Strain Diversity of *Mycobacterium tuberculosis* Isolates from Pulmonary Tuberculosis Patients in Afar Pastoral Region of Ethiopia. BioMed Research International *2014*, e238532.

Bravo, L.T.C., and Procop, G.W. (2009). Recent advances in diagnostic microbiology. Semin. Hematol. *46*, 248–258.

Brites, D., and Gagneux, S. (2015). Co-evolution of *Mycobacterium tuberculosis* and Homo sapiens. Immunol. Rev. *264*, 6–24.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., et al. (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc. Natl. Acad. Sci. U.S.A. *99*, 3684–3689.

Brown, T., Nikolayevskyy, V., Velji, P., and Drobniewski, F. (2010). Associations between *Mycobacterium tuberculosis* strains and phenotypes. Emerging Infect. Dis *16*, 272–280.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., et al. (2006). BMC Microbiol. *6*, 23.

Cavalcanti, Y.V.N., Brelaz, M.C.A., Neves, J.K. de A.L., Ferraz, J., Candido, Pereira, V., R&#xea, R., and Alves, G. (2012). Role of TNF-Alpha, IFN-Gamma, and IL-10 in the Development of Pulmonary Tuberculosis. Pulmonary Medicine *2012*, e745483.

Chang, J.C., Miner, M.D., Pandey, A.K., Gill, W.P., Harik, N.S., Sassetti, C.M., and Sherman, D.R. (2009). igr Genes and *Mycobacterium tuberculosis* cholesterol metabolism. J. Bacteriol *191*, 5232–5239.

Chatterjee, A., D'Souza, D., Vira, T., Bamne, A., Ambe, G.T., Nicol, M.P., Wilkinson, R.J., and Mistry, N. (2010). Strains of *Mycobacterium tuberculosis* from western Maharashtra, India, exhibit a high degree of diversity and strain-specific associations with drug resistance, cavitary disease, and treatment failure. J. Clin. Microbiol *48*, 3593–3599.

Chauhan, S., and Tyagi, J.S. (2011). Analysis of transcription at the oriC locus in *Mycobacterium tuberculosis* . Microbiol. Res. *166*, 508–514.

Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: 99. In Proceedings of the German Conference on Bioinformatics (GCB), pp. 45–56.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E.G., Wetter, T., and

Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. *14*, 1147–1159.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth *10*, 563–569.

Chuang, P.-C., Chen, Y.-M.A., Chen, H.-Y., and Jou, R. (2010a). Single nucleotide polymorphisms in cell wall biosynthesis-associated genes and phylogeny of *Mycobacterium tuberculosis* lineages. Infect. Genet. Evol *10*, 459–466.

Chuang, P.-C., Chen, H.-Y., and Jou, R. (2010b). Single-nucleotide polymorphism in the fadD28 gene as a genetic marker for East Asia Lineage *Mycobacterium tuberculosis* . J. Clin. Microbiol *48*, 4245–4247.

Coll, F., McNerney, R., Guerra-Assunção, J.A., Glynn, J.R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N., and Clark, T.G. (2014). A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun *5*, 4812.

Comas, I., Homolka, S., Niemann, S., and Gagneux, S. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS ONE *4*, e7815.

Comas, I., Chakravartti, J., Small, P.M., Galagan, J., Niemann, S., Kremer, K., Ernst, J.D., and Gagneux, S. (2010). Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet *42*, 498–503.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., et al. (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet *advance online publication*.

Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R., and Young, D.B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis* . Cell Rep *5*, 1121–1131.

Cowley, D., Govender, D., February, B., Wolfe, M., Steyn, L., Evans, J., Wilkinson, R.J., and Nicol, M.P. (2008). Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. Clin. Infect. Dis *47*, 1252–1259.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics *11*, 485.

Dambuza, I., Allie, N., Fick, L., Johnston, N., Fremond, C., Mitchell, J., Quesniaux, V.F.J., Ryffel, B., and Jacobs, M. (2008). Efficacy of membrane TNF mediated host resistance is dependent on mycobacterial virulence. Tuberculosis (Edinb) *88*, 221–234.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLoS ONE *5*, e11147.

Demay, C., Liens, B., Burguière, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., and Rastogi, N. (2012). SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infection, Genetics and Evolution *12*, 755–766.

Denisov, G., Walenz, B., Halpern, A.L., Miller, J., Axelrod, N., Levy, S., and Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. Bioinformatics *24*, 1035–1040.

Domenech, P., Kolly, G.S., Leon-Solis, L., Fallow, A., and Reed, M.B. (2010). Massive gene duplication event among clinical isolates of the *Mycobacterium tuberculosis* W/Beijing family. J. Bacteriol *192*, 4562–4570.

Dos Vultos, T., Mestre, O., Rauzier, J., Golec, M., Rastogi, N., Rasolofo, V., Tonjum, T., Sola, C., Matic, I., and Gicquel, B. Evolution and Diversity of Clonal Bacteria: The Paradigm of *Mycobacterium tuberculosis* . PLoS ONE *3*.

Ebrahimi-Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., Kreiswirth, B., Blazquez, J., Jouan, M., van Soolingen, D., et al. (2003). Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. Emerging Infect. Dis *9*, 838–845.

van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., and Shinnick, T.M. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. *31*, 406–409.

Ferdinand, S., Sola, C., Chanteau, S., Ramarokoto, H., Rasolonavalona, T., Rasolofo-Razanamparany, V., and Rastogi, N. (2005). A study of spoligotyping-defined *Mycobacterium tuberculosis* clades in relation to the origin of peopling and the demographic history in Madagascar. Infect. Genet. Evol *5*, 340–348.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbón, M.H., Bobadilla del Valle, M., Fyfe, J., García-García, L., Rastogi, N., Sola, C., et al. (2006). Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol *188*, 759–772.

Flores, L., Van, T., Narayanan, S., DeRiemer, K., Kato-Maeda, M., and Gagneux, S. (2007). Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. J. Clin. Microbiol *45*, 3393–3395.

Forse, L.N., Houghton, J., and Davis, E.O. (2011). Enhanced expression of recX in *Mycobacterium tuberculosis* owing to a promoter internal to recA. Tuberculosis (Edinb) *91*, 127–135.

Gagneux, S. (2012). Host-pathogen coevolution in human tuberculosis. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *367*, 850–859.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., et al. (2006). Variable host-pathogen compatibility in *Mycobacterium tuberculosis* . Proc. Natl. Acad. Sci. U.S.A. *103*, 2869–2873.

Galagan, J.E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., et al. (2013). The *Mycobacterium tuberculosis* regulatory network and hypoxia. Nature *499*, 178–183.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., and Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics *28*, 2678–2679.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol *27*, 221–224.

Griffin, J.E., Gawronski, J.D., Dejesus, M.A., Ioerger, T.R., Akerley, B.J., and Sassetti, C.M. (2011). High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog. *7*, e1002251.

Gutacker, M.M., Mathema, B., Soini, H., Shashkina, E., Kreiswirth, B.N., Graviss, E.A., and Musser, J.M. (2006). Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J. Infect. Dis. *193*, 121–128.

Gutierrez, M.C., Brisse, S., Brosch, R., Fabre, M., Omaïs, B., Marmiesse, M., Supply, P., and Vincent, V. (2005). Ancient Origin and Gene Mosaicism of the Progenitor of *Mycobacterium tuberculosis* . PLoS Pathog *1*, e5.

Gutierrez, M.C., Ahmed, N., Willery, E., Narayanan, S., Hasnain, S.E., Chauhan, D.S., Katoch, V.M., Vincent, V., Locht, C., and Supply, P. (2006). Predominance of ancestral lineages of *Mycobacterium tuberculosis* in India. Emerging Infect. Dis *12*, 1367–1374.

Hanekom, M., van der Spuy, G.D., Streicher, E., Ndabambi, S.L., McEvoy, C.R.E., Kidd, M., Beyers, N., Victor, T.C., van Helden, P.D., and Warren, R.M. (2007a). A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. J. Clin. Microbiol *45*, 1483–1490.

Hanekom, M., van der Spuy, G.D., Streicher, E., Ndabambi, S.L., McEvoy, C.R.E., Kidd, M., Beyers, N., Victor, T.C., van Helden, P.D., and Warren, R.M. (2007b). A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. J. Clin. Microbiol. *45*, 1483–1490.

Hanekom, M., van der Spuy, G.D., van Pittius, N.C.G., McEvoy, C.R.E., Ndabambi, S.L., Victor, T.C., Hoal, E.G., van Helden, P.D., and Warren, R.M. (2007c). Evidence that the Spread of *Mycobacterium tuberculosis* Strains with the Beijing Genotype Is Human Population Dependent. J. Clin. Microbiol. *45*, 2263–2266.

Hannes Ponstigl (2014). SMALT (Welcome Trust Sanger Institute).

Hatem, A., Bozdağ, D., Toland, A.E., and Çatalyürek, Ü.V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics *14*, 184.

Helal, Z.H., Ashour, M.S.E.-D., Eissa, S.A., Abd-Elatef, G., Zozio, T., Babapoor, S., Rastogi, N., and Khan, M.I. (2009). Unexpectedly high proportion of ancestral Manu genotype *Mycobacterium tuberculosis* strains cultured from tuberculosis patients in Egypt. J. Clin. Microbiol. *47*, 2794–2801.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., et al. (2008). High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. *6*, e311.

Homer, N., Merriman, B., and Nelson, S.F. (2009). BFAST: An Alignment Tool for Large Scale Genome Resequencing. PLoS One *4*.

Homolka, S., Niemann, S., Russell, D.G., and Rohde, K.H. (2010). Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog. *6*, e1000988.

Jagielski, T., van Ingen, J., Rastogi, N., Dziadek, J., aw, Mazur, P., K, and Bielecki, J. (2014). Current Methods in the Molecular Typing of *Mycobacterium tuberculosis* and Other Mycobacteria. BioMed Research International *2014*, e645802.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. gr.170720.113.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., et al. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. *35*, 907–914.

Kanji, A., Hasan, Z., Tanveer, M., Laiq, R., and Hasan, R. (2011a). Occurrence of RD149 and RD152 deletions in *Mycobacterium tuberculosis* strains from Pakistan. J Infect Dev Ctries *5*, 106–113.

Kanji, A., Hasan, Z., Tanveer, M., Mahboob, R., Jafri, S., and Hasan, R. (2011b). Presence of RD149 deletions in M. tuberculosis Central Asian Strain 1 isolates affect growth and TNFα induction in THP-1 monocytes. PLoS ONE *6*, e24178.

Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., and Mclnerney, J.O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology *6*, 29.

Kibiki, G.S., Mulder, B., Dolmans, W.M.V., de Beer, J.L., Boeree, M., Sam, N., van Soolingen, D., Sola, C., and van der Zanden, A.G.M. (2007). M. tuberculosis

genotypic diversity and drug susceptibility pattern in HIV-infected and non-HIV-infected patients in northern Tanzania. BMC Microbiol *7*, 51.

Klopper, M., Warren, R.M., Hayes, C., Gey van Pittius, N.C., Streicher, E.M., Müller, B., Sirgel, F.A., Chabula-Nxiweni, M., Hoosain, E., Coetzee, G., et al. (2013). Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa. Emerging Infect. Dis. *19*, 449–455.

Koren, S., and Phillippy, A.M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology *23*, 110–120.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. *30*, 693–700.

Kruh, N.A., Troudt, J., Izzo, A., Prenni, J., and Dobos, K.M. (2010). Portrait of a pathogen: the *Mycobacterium tuberculosis* proteome in vivo. PLoS ONE *5*, e13938.

Kumar, D., Nath, L., Kamal, M.A., Varshney, A., Jain, A., Singh, S., and Rao, K.V.S. (2010). Genome-wide Analysis of the Host Intracellular Network that Regulates Survival of *Mycobacterium tuberculosis* . Cell *140*, 731–743.

Lee, J.S., Krause, R., Schreiber, J., Mollenkopf, H.-J., Kowall, J., Stein, R., Jeon, B.-Y., Kwak, J.-Y., Song, M.-K., Patron, J.P., et al. (2008). Mutation in the transcriptional regulator PhoP contributes to avirulence of *Mycobacterium tuberculosis* H37Ra strain. Cell Host Microbe *3*, 97–103.

Lew, J.M., Kapopoulou, A., Jones, L.M., and Cole, S.T. (2011). TubercuList – 10 years after. Tuberculosis *91*, 1–7.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y.-C., Lin, S.-H., and Lin, H.-H. (2015). Completing bacterial genome assemblies: strategy and performance comparisons. Sci. Rep. *5*.

Liu, F., Hu, Y., Wang, Q., Li, H.M., Gao, G.F., Liu, C.H., and Zhu, B. (2014). Comparative genomic analysis of *Mycobacterium tuberculosis* clinical isolates. BMC Genomics *15*, 469.

Luo, T., Comas, I., Luo, D., Lu, B., Wu, J., Wei, L., Yang, C., Liu, Q., Gan, M., Sun, G., et al. (2015). Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. Proc. Natl. Acad. Sci. U.S.A. *112*, 8136–8141.

Maddison W.P. Maddison D.R. (2015). Mesquite: a modular system for evolutionary analysis.

Malen, H., De Souza, G., Pathak, S., Softeland, T., and Wiker, H. (2011). Comparison of membrane proteins of *Mycobacterium tuberculosis* H37Rv and H37Ra strains. BMC Microbiology *11*, 18.

Mathuria, J.P., Sharma, P., Prakash, P., Samaria, J.K., Katoch, V.M., and Anupurba, S. (2008). Role of spoligotyping and IS6110-RFLP in assessing genetic diversity of *Mycobacterium tuberculosis* in India. Infect. Genet. Evol *8*, 346–351.

Mendum, T.A., Wu, H., Kierzek, A.M., and Stewart, G.R. (2015). Lipid metabolism and Type VII secretion systems dominate the genome scale virulence profile of *Mycobacterium tuberculosis* in human dendritic cells. BMC Genomics *16*, 372.

Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., Blum, M.G.B., Rüsch-Gerdes, S., Mokrousov, I., Aleksic, E., et al. (2015). Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat. Genet. *47*, 242–249.

Mestre, O., Luo, T., Dos Vultos, T., Kremer, K., Murray, A., Namouchi, A., Jackson, C., Rauzier, J., Bifani, P., Warren, R., et al. (2011). Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS ONE *6*, e16020.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat Protoc *8*, 1551–1566.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24*, 2818–2824.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. Genomics *95*, 315–327.

Millet, J., Miyagi-Shiohira, C., Yamane, N., Mokrousov, I., and Rastogi, N. (2012). High-resolution MIRU-VNTRs typing reveals the unique nature of *Mycobacterium tuberculosis* Beijing genotype in Okinawa, Japan. Infect. Genet. Evol. *12*, 637–641.

Minch, K.J., Rustad, T.R., Peterson, E.J.R., Winkler, J., Reiss, D.J., Ma, S., Hickey, M., Brabant, W., Morrison, B., Turkarslan, S., et al. (2015). The DNA-binding network of *Mycobacterium tuberculosis* . Nat Commun *6*, 5829.

Mokrousov, I., Narvskaya, O., Otten, T., Vyazovaya, A., Limeschenko, E., Steklova, L., and Vyshnevskyi, B. (2002). Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. Res. Microbiol. *153*, 629–637.

Narayanan, P.R. (2006). Influence of sex, age & nontuberculous infection at intake on the efficacy of BCG: re-analysis of 15-year data from a double-blind randomized control trial in South India. Indian J. Med. Res. *123*, 119–124.

Narayanan, S., Gagneux, S., Hari, L., Tsolaki, A.G., Rajasekhar, S., Narayanan, P.R., Small, P.M., Holmes, S., and Deriemer, K. (2008). Genomic interrogation of ancestral *Mycobacterium tuberculosis* from south India. Infect. Genet. Evol. *8*, 474–483.

Newton, S.M., Smith, R.J., Wilkinson, K.A., Nicol, M.P., Garton, N.J., Staples, K.J., Stewart, G.R., Wain, J.R., Martineau, A.R., Fandrich, S., et al. (2006). A deletion defining a common Asian lineage of *Mycobacterium tuberculosis* associates with immune subversion. Proc. Natl. Acad. Sci. U.S.A. *103*, 15594–15598.

Newton-Foot, M., and Gey van Pittius, N.C. (2013). The complex architecture of mycobacterial promoters. Tuberculosis *93*, 60–74.

Portevin, D., Gagneux, S., Comas, I., and Young, D. (2011). Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog. *7*, e1001307.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 341.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics *47*, 11.12.1–11.12.34.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rakotosamimanana, N., Raharimanga, V., Andriamandimby, S.F., Soares, J.-L., Doherty, T.M., Ratsitorahina, M., Ramarokoto, H., Zumla, A., Huggett, J., Rook, G., et al. (2010). Variation in gamma interferon responses to different infecting strains of *Mycobacterium tuberculosis* in acid-fast bacillus smear-positive patients and household contacts in Antananarivo, Madagascar. Clin. Vaccine Immunol *17*, 1094–1103.

Redford, P.S., Murray, P.J., and O'Garra, A. (2011). The role of IL-10 in immune regulation during M. tuberculosis infection. Mucosal Immunol *4*, 261–270.

Reed, M.B., Pichler, V.K., McIntosh, F., Mattia, A., Fallow, A., Masala, S., Domenech, P., Zwerling, A., Thibert, L., Menzies, D., et al. (2009). Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. J. Clin. Microbiol. *47*, 1119–1128.

Rengarajan, J., Bloom, B.R., and Rubin, E.J. (2005). Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. Proc. Natl. Acad. Sci. U.S.A. *102*, 8327–8332.

Roetzer, A., Diel, R., Kohl, T.A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., et al. (2013). Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med. *10*, e1001387.

Rosas-Magallanes, V., Deschavanne, P., Quintana-Murci, L., Brosch, R., Gicquel, B.,

and Neyrolles, O. (2006). Horizontal Transfer of a Virulence Operon to the Ancestor of *Mycobacterium tuberculosis* . Mol Biol Evol *23*, 1129–1135.

Rose, G., Cortes, T., Comas, I., Coscolla, M., Gagneux, S., and Young, D.B. (2013). Mapping of genotype-phenotype diversity among clinical isolates of *Mycobacterium tuberculosis* by sequence-based transcriptional profiling. Genome Biol Evol *5*, 1849–1862.

Ruffalo, M., LaFramboise, T., and Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics *27*, 2790–2796.

Salie, M., van der Merwe, L., Möller, M., Daya, M., van der Spuy, G.D., van Helden, P.D., Martin, M.P., Gao, X.-J., Warren, R.M., Carrington, M., et al. (2014). Associations between human leukocyte antigen class I variants and the *Mycobacterium tuberculosis* subtypes causing disease. J. Infect. Dis. *209*, 216–223.

Schürch, A.C., Kremer, K., Hendriks, A.C.A., Freyee, B., McEvoy, C.R.E., van Crevel, R., Boeree, M.J., van Helden, P., Warren, R.M., Siezen, R.J., et al. (2011a). SNP/RD typing of *Mycobacterium tuberculosis* Beijing strains reveals local and worldwide disseminated clonal complexes. PLoS ONE *6*, e28365.

Schürch, A.C., Kremer, K., Warren, R.M., Hung, N.V., Zhao, Y., Wan, K., Boeree, M.J., Siezen, R.J., Smith, N.H., and van Soolingen, D. (2011b). Mutations in the regulatory network underlie the recent clonal expansion of a dominant subclone of the *Mycobacterium tuberculosis* Beijing genotype. Infect. Genet. Evol *11*, 587–597.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int *2014*, 309650.

Shitikov, E.A., Bespyatykh, J.A., Ischenko, D.S., Alexeev, D.G., Karpova, I.Y., Kostryukova, E.S., Isaeva, Y.D., Nosova, E.Y., Mokrousov, I.V., Vyazovaya, A.A., et al. (2014). Unusual Large-Scale Chromosomal Rearrangements in *Mycobacterium tuberculosis* Beijing B0/W148 Cluster Isolates. PLoS ONE *9*, e84971.

Singh, U.B., Suresh, N., Bhanu, N.V., Arora, J., Pant, H., Sinha, S., Aggarwal, R.C., Singh, S., Pande, J.N., Sola, C., et al. (2004). Predominant tuberculosis spoligotypes, Delhi, India. Emerging Infect. Dis *10*, 1138–1142.

Singh, U.B., Arora, J., Suresh, N., Pant, H., Rana, T., Sola, C., Rastogi, N., and Pande, J.N. (2007). Genetic biodiversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in India. Infection, Genetics and Evolution *7*, 441–448.

Sola, C., Filliol, I., Legrand, E., Lesjean, S., Locht, C., Supply, P., and Rastogi, N. (2003). Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect. Genet. Evol *3*, 125–133.

van Soolingen, D., and Kremer, K. (2009). [Findings and ongoing research in the molecular epidemiology of tuberculosis]. Kekkaku *84*, 83–89.

de Souza, G.A., and Wiker, H.G. (2011). A proteomic view of mycobacteria. Proteomics *11*, 3118–3127.

de Souza, G.A., Fortuin, S., Aguilar, D., Pando, R.H., McEvoy, C.R.E., van Helden, P.D., Koehler, C.J., Thiede, B., Warren, R.M., and Wiker, H.G. (2010). Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hypervirulent clinical *Mycobacterium tuberculosis* Beijing isolates. Mol. Cell Proteomics *9*, 2414–2423.

de Souza, G.A., Leversen, N.A., Målen, H., and Wiker, H.G. (2011). Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway. Journal of Proteomics.

van der Spuy, G.D., Kremer, K., Ndabambi, S.L., Beyers, N., Dunbar, R., Marais, B.J., van Helden, P.D., and Warren, R.M. (2009). Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. Tuberculosis (Edinb) *89*, 120–125.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., and Musser, J.M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. U.S.A. *94*, 9869–9874.

Streicher, E.M., Victor, T.C., van der Spuy, G., Sola, C., Rastogi, N., van Helden, P.D., and Warren, R.M. (2007). Spoligotype signatures in the *Mycobacterium tuberculosis* complex. J. Clin. Microbiol. *45*, 237–240.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rüsch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., et al. (2006). Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis* . J. Clin. Microbiol. *44*, 4498–4510.

Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M., and Otto, T.D. (2012). A Post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc *7*, 1260–1284.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.

Thomas, S.K., Iravatham, C.C., Moni, B.H., Kumar, A., Archana, B.V., Majid, M., Priyadarshini, Y., Rani, P.S., Valluri, V., Hasnain, S.E., et al. (2011). Modern and Ancestral Genotypes of *Mycobacterium tuberculosis* from Andhra Pradesh, India. PLoS ONE *6*, e27584.

Tsolaki, A.G., Hirsh, A.E., DeRiemer, K., Enciso, J.A., Wong, M.Z., Hannan, M., de la Salmoniere, Y.-O.L.G., Aman, K., Kato-Maeda, M., and Small, P.M. (2004). Functional and evolutionary genomics of *Mycobacterium tuberculosis* : Insights from genomic deletions in 100 strains. Proceedings of the National Academy of Sciences

of the United States of America *101*, 4865–4870.

Tsolaki, A.G., Gagneux, S., Pym, A.S., Goguet de la Salmoniere, Y.-O.L., Kreiswirth, B.N., Van Soolingen, D., and Small, P.M. (2005). Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis* . J. Clin. Microbiol *43*, 3185–3191.

Turner, F.S. (2014). Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. Front Genet *5*, 5.

Viegas, S.O., Machado, A., Groenheit, R., Ghebremichael, S., Pennhag, A., Gudo, P.S., Cuna, Z., Miotto, P., Hill, V., Marrufo, T., et al. (2010). Molecular diversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mozambique. BMC Microbiol *10*, 195.

Wada, T., Iwamoto, T., and Maeda, S. (2009). Genetic diversity of the *Mycobacterium tuberculosis* Beijing family in East Asia revealed through refined population structure analysis. FEMS Microbiology Letters *291*, 35–43.

Wang, C., Peyron, P., Mestre, O., Kaplan, G., van Soolingen, D., Gao, Q., Gicquel, B., and Neyrolles, O. (2010). Innate immune response to *Mycobacterium tuberculosis* Beijing and other genotypes. PLoS ONE *5*, e13594.

Warren, R., Richardson, M., van der Spuy, G., Victor, T., Sampson, S., Beyers, N., and van Helden, P. (1999). DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. Electrophoresis *20*, 1807–1812.

Warren, R.M., Victor, T.C., Streicher, E.M., Richardson, M., Beyers, N., Gey van Pittius, N.C., and van Helden, P.D. (2004). Patients with active tuberculosis often have different strains in the same sputum specimen. Am. J. Respir. Crit. Care Med *169*, 610–614.

Weiner, B., Gomez, J., Victor, T.C., Warren, R.M., Sloutsky, A., Plikaytis, B.B., Posey, J.E., van Helden, P.D., Gey van Pittius, N.C., Koehrsen, M., et al. (2012). Independent large scale duplications in multiple M. tuberculosis lineages overlapping the same genomic region. PLoS ONE *7*, e26038.

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., et al. (2008). Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog. *4*, e1000160.

SUPPLEMENTAL DATA

(see attached CD)